Social Network Analysis and Mining (2024) 14:53
https://doi.org/10.1007/s13278-024-01201-4

ORIGINAL ARTICLE

A two-stage framework for Arabic social media text misinformation detection combining data augmentation and AraBERT

Ebtsam A. Mohamed[1] · Walaa N. Ismail[1,2] · Osman Ali Sadek Ibrahim[3] · Eman M. G. Younis[1]

Received: 23 November 2023 / Revised: 31 December 2023 / Accepted: 8 January 2024

Abstract

Misinformation can profoundly impact the reputation of an entity, and eliminating its spread has become a critical concern across various applications. Social media, often a primary source of information, can significantly influence individuals' perspectives through content from less credible sources. The utilization of machine-learning (ML) algorithms can facilitate automated, large-scale analysis of textual content, contributing to the rapid and efficient processing of extensive datasets for informed decision-making. Since the performance of ML models is highly affected by the size of the training data, many research papers have presented different approaches to solve the problem of limited dataset size. The data augmentation (DA) approach is one of these strategies, aiming to enhance ML model performance by increasing the amount of training data. DA generates new instances by applying different transformations to the original data instances. While many DA techniques have been investigated for various languages, such as English, achieving an enhancement of the classification model's performance on the new augmented dataset compared to the original dataset, there is a lack of studies on the Arabic language due to its unique characteristics. This paper introduces a novel two-stage framework designed for the automated identification of misinformation in Arabic textual content. The first stage aims to identify the optimal representation of features before feeding them to the ML model. Diverse representations of tweet content are explored, including *N*-grams, content-based features, and source-based features. The second stage focuses on investigating the DA effect through the back-translation technique applied to the original training data. Back-translation entails translating sentences from the target language (in this case, Arabic) into another language and then back to Arabic. As a result of this procedure, new examples for training are created by introducing variances in the text. The study utilizes support vector machine (SVM), naive Bayes, logistic regression (LR), and random forest (RF) as baseline algorithms. Additionally, AraBERT transformer pre-trained language models are used to relate the instance's label and feature representation of the input. Experimental outcomes demonstrate that misinformation detection, coupled with data augmentation, enhances accuracy by a noteworthy margin 5 to 12% compared to baseline machine-learning algorithms and pre-trained models. Remarkably, the results show the superiority of the *N*-grams approach over traditional state-of-the-art feature representations concerning accuracy, recall, precision, and *F*-measure metrics. This suggests a promising avenue for improving the efficacy of misinformation detection mechanisms in the realm of Arabic text analysis.

1 Introduction

A social media platform such as Twitter holds considerable influence over people's attitudes during emergencies (Cuesta et al. 2013). Twitter, being a swift and accessible medium, has evolved into a primary channel for disseminating news and updates among family, friends, and the public, outpacing traditional media in information distribution. However, the prevalence of misinformation during crises, often originating from non-experts and community members rather than certified specialists, underscores the imperative to assess the credibility of tweets (Mourad

✉ Ebtsam A. Mohamed
ebtesam.mohamed@mu.edu.eg

Walaa N. Ismail
W_abdelfattah@yu.edu.sa

Eman M. G. Younis
eman.younas@mu.edu.eg

[1] Faculty of Computers and Information, Minia University, Minya, Egypt

[2] College of Business Administration, Al Yamamah University, Riyadh, Saudi Arabia

[3] Faculty of Science, Minia University, Minya, Egypt

Published online: 08 March 2024

et al. 2020). Research studies consistently reveal the non-credibility of Twitter content, with findings indicating that a substantial portion, approximately 40%, of daily tweets lack credibility (El Ballouli et al. 2017). This platform sometimes becomes a dumping ground for false information and rumors (Lu and Brelsford 2014). Moreover, there is a growing need for efforts to identify tweets written in multiple languages (Mohamed et al. 2019b, a). Categorizing Arabic tweets has been the subject point of numerous attempts, often employing machine-learning techniques that consider both content-related and source-related features (Gupta et al. 2023; Yadav et al. 2023; Asta and Setiawan 2023). Content-related features encompass tweet length and hashtag usage, while source-related features, include user verification and followers' count (El Ballouli et al. 2017; Hassan et al. 2018). Although several recent studies have explored content and source features for misinformation detection, manual feature creation remains challenging, time-consuming, and may not be accurate enough for classification (El Ballouli et al. 2017; Hassan et al. 2018; Capuano et al. 2023). In addition, certain features (such as followers' number), often used as a measure of credibility, may be misleading due to their changeability, as well as the tendency of users to reshare posts without verifying their accuracy (Ravikumar et al. 2012).

Unlike English, Arabic possesses a complex morphological structure in which words may undergo morphological changes depending on the context, tense, gender, and other linguistic factors (Gupta et al. 2023; Yadav et al. 2023; Hassan et al. 2018). This morphological complexity can be effectively captured through the use of contextual embeddings. Employing contextual embeddings addresses static word context challenges by capturing the context preceding and following each word. This technique proves particularly useful in identifying misinformation, requiring no additional features beyond the text itself and capturing word context as phrases, thereby enhancing the detection of misinformation. Additionally, *N*-grams serve as an effective method for capturing the linguistic nuances and contextual complexities inherent in Arabic text classification (Kumar et al. 2020; Hua et al. 2023). In this method, contiguous sequences of *N* words in a given text are represented, providing a reliable approach for feature extraction without necessitating manual preparation of features. By encoding word context as phrases, *N*-grams offer a valuable solution in the Arabic literature, where intricate morphology makes tokenization and stemming challenging. A further challenge in the classification of Arabic text is the lack of benchmark datasets. Data augmentation (DA) methods are proposed for generating synthetic data, alleviating the constraints of limited labeled data (Jiang et al. 2023; Wang et al. 2023). To overcome the challenges mentioned above, the contributions of this research are outlined as follows:

1. Introduction of a novel two-stage framework for detecting misinformation in Arabic tweets, integrating machine-learning algorithms and data augmentation (DA) to enhance the performance of the misinformation classification task.
2. Exploring various feature representations of input text to aid the classification model in accurately predicting labels. The classification model is evaluated using *N*-gram, content-based, and source-based features with publicly available datasets ('news,' 'covid1,' and 'covid2').
3. Utilizing contextual embeddings through AraBERT to address the limitations of static word embeddings, which produce a single representation for words with different meanings, thereby enabling the model to capture nuances in meaning based on the context.
4. Building and fine-tuning the hyperparameters of four baseline machine-learning algorithms and investigating their performance on two standard datasets for Arabic tweets. This comprehensive assessment provides insights into their effectiveness in handling diverse linguistic contexts.
5. Implementing the DA back-translation technique to augment the training data and preserve the data distribution of the original dataset. The application of DA results in substantial performance enhancements, including a significant 12% increase in overall *F*-measure and a notable 13% improvement in recall. These improvements underscore the efficacy of the proposed framework in accurately detecting and classifying misinformation in Arabic tweets, even in scenarios with a limited number of labeled tweets.

2 Background and related work

2.1 Data augmentation for text classification

The goal of data augmentation techniques is to enhance data quality and diversity without the need for additional data collection. Several fields have successfully applied data augmentation (DA) to improve model performance and generalization (Jiang et al. 2023; Wang et al. 2023; Yadav and Vishwakarma 2023). To accomplish this, new samples are generated through methods such as noise addition, cropping, or flipping the training dataset while maintaining the integrity of the original dataset (Li et al. 2022, 2023; Al-Dhabyani et al. 2019).

Unlike computer vision, which focuses on pixels, NLP data augmentation emphasizes linguistic variations. Utilized methods include synonym substitution, sentence shuffling, back-translation, and embedding modification. These solutions encompass tasks such as balancing classes

within unbalanced datasets and generating additional data for under-resourced domains (Wang et al. 2023; Bayer et al. 2023). Text DA can be implemented based on feature space or data space (Bayer et al. 2022). In data space DA, adjustments can be made at various granularities, including the character level, word level, sentence level, or document level (Bayer et al. 2022). These strategies aim to expose the model to a broad array of linguistic patterns, fostering improved generalization and performance across diverse NLP tasks such as sentiment analysis and named entity recognition (Wang et al. 2023; Jiang et al. 2023). Deep learning models, including transfer models like BERT, have gained popularity in natural language processing (NLP) tasks owing to their ability to streamline feature extraction efforts (Gupta et al. 2023; Yadav et al. 2023; Asta and Setiawan 2023). While the use of data augmentation in NLP is still in its early stages, it plays a crucial role in addressing the lack of data and strengthening the resilience of language models as the field continues to evolve (Pellicer et al. 2023; Wang et al. 2023).

Text augmentation for fake news detection is becoming increasingly popular as a method for improving model resilience and generalization when dealing with the growing number of fake news sources (Refai et al. 2022). A variety of methods are employed to augment or expand the training dataset, enhancing the model's capacity to handle a broader range of cases that simulate the complexity and diversity found in real-life misleading information (Refai et al. 2022; Kumar et al. 2020; Hua et al. 2023). These methods include back-translation, paraphrasing, and context-based augmentation. Back translation is a method that utilizes translation models to create paraphrases. In this approach, text is translated into another language and then back-translated into the original language. The underlying concept of back-translation is rooted in the complexity of natural language processing (NLP), which results in multiple translations for a given text. This approach is highly effective due to its high paraphrasing capabilities and the preservation of labels for newly generated instances. Back-translation outperformed the EDA method, GPT2, and BERT pre-trained models for data augmentation (DA) (Kumar et al. 2020; Hua et al. 2023). A study conducted by (Kumar et al. 2020), the authors applied DA based on back-translation to two datasets. Their experimental findings demonstrated that the back-translation technique surpasses other methods in terms of precision, thus demonstrating the efficiency of contemporary translation systems in maintaining language semantics. In the work presented by Refai et al. (2022), a novel DA method was introduced for Arabic text classification, aiming to integrate the unique features of the Arabic language. The motivation behind this endeavor stemmed from the established efficacy of textual augmentation in enhancing the performance of text classification tasks. The authors utilized Arabic transformers, specifically AraGPT-2 and AraBERT, for the generation and processing of Arabic text. Furthermore, they employed well-known similarity metrics, including cosine, Euclidean, Jaccard, and BLEU measures, to ensure the quality of the augmented text. This consideration encompassed aspects such as diversity, context, and semantics. Additionally, in Sabty et al. (2021), data augmentation (DA) was employed to augment the limited amount of labeled data available for named entity recognition (NER). Various automatic augmentation techniques, such as back-translation, modified EDA, and word embedding substitution, were introduced to expand the training data and enhance the performance of Arabic NER. The study's findings demonstrated that the performance of NER can be improved through the application of combinations of different DA techniques.

2.2 Misinformation detection in Arabic language

Automatically predicting misleading information in Arabic social media is both a technological and socio-cultural imperative. Addressing inaccurate information on Arabic social media is crucial for several reasons (Albalawi et al. 2023; Singh et al. 2023). Firstly, misinformation can significantly impact public perceptions, potentially giving rise to misguided beliefs, fostering fear, or even inciting unjustified actions. Secondly, the linguistic and cultural characteristics of Arabic pose specific challenges for the automated identification of misleading information, necessitating specialized methodologies and models.

In social media platforms, machine-learning (ML) algorithms offer a sophisticated and efficient way to analyze large volumes of textual and contextual data. Recent research in fake news detection sheds light on two primary approaches: classification and propagation (El Ballouli et al. 2017; Jin et al. 2014; Albalawi et al. 2023). Propagation-based approaches (Singh et al. 2023; Azad 2023) delve into the analysis of social graph structures to identify misinformation (Jin et al. 2014). In contrast, classification-based approaches employ machine-learning algorithms that rely on textual features extracted from the content itself (El Ballouli et al. 2017; Zubiaga et al. 2017; Sabbeh and BAATWAH 2018).

Additionally, advancements in the field have introduced two key dimensions to misinformation detection: source-based and content-based features. Content-based approaches rely on factors such as text length, the presence of hashtags (#) in the text, and sentiment features (El Ballouli et al. 2017; Kazmi et al. 2023). On the other hand, source-based features are derived from user characteristics, including follower count and user account verification. Some studies, such as (El Ballouli et al. 2017; Hassan et al. 2018), propose a hybrid approach, combining both source and content features for a more comprehensive analysis.

Lorek et al. (2015) automated the matching between the contents of external links and text content as an evidence

feature for classifying tweets as credible or incredible. Zubiaga et al. (2017) introduced a misinformation detection system to help users identify fake tweets using the conditional random fields algorithm. Their approach is based on content-based features and social features to compare results. According to the obtained results, the textual features of the tweet's content effectively detect the tweet's credibility.

Furthermore, Hassan et al. (2018) examined different feature sets, including content and source features, over the dataset used in (Lorek et al. 2015). Their work concluded that using features related to the source is better than using features related to the content. Results showed that source features improved *F*-measure by 49% (Lorek et al. 2015). Unfortunately, the creation of handcrafted features is not only time-consuming but also poses the risk of being misleading. An illustrative example is the reliance on the number of followers or reshares as indicators of a tweet's credibility, a metric that may not reliably signify the verification of content by users before resharing (Ravikumar et al. 2012).

In the domain of Arabic misinformation, a considerable challenge arises from the scarcity of labeled datasets essential for training machine-learning (ML) algorithms. Recent research efforts in the field have explored deep learning (DL) approaches to tackle misinformation detection in Arabic blogs. In (Gaanoun 2020), the authors employed a semi-supervised technique based on the Arabic-BERT model and ensemble models after the DA process. Their approach achieved higher performance using Arabic-BERT models compared to baseline models. However, the majority of DA works for Arabic focus on sentiment analysis tasks. In Abuzayed and Al-Khalifa (2021), several BERT models were utilized, and the DA process was applied to tweets to enhance the performance of sentiment detection. Another study (Alkadri 2022) investigated the effect of using DA for Arabic tweets in the spam detection task. Their focus was on addressing the dataset imbalance problem by increasing the instances of minority classes. They utilized a large corpus to extract Word2Vec embedding vectors to represent tweet contents. A notable exploration into DL methodologies was conducted by Ajao et al. (2018), employing recurrent neural networks (RNN) and long short-term memory models (LSTM) for social media text classification. This research achieved an impressive 82% accuracy using the dataset also utilized in previous works (Hassan et al. 2018; Lorek et al. 2015). The advantage of DL models lies in their automatic feature extraction, eliminating the need for manual crafting of features. However, the substantial requirement for labeled data in training remains a drawback.

In another direction, a novel approach was introduced using an *N*-gram model for misinformation detection in tweets, as proposed by Hassan et al. (2020). *N*-gram features, illuminating word relationships and their context within a sentence, were applied to both the Arabic dataset from El Ballouli et al. (2017) and the PHEME dataset. Remarkably, the *N*-grams model exhibited superior performance on the PHEME and CAT datasets compared to the LSTM DL model presented by Ajao et al. (2018) for fake tweet detection. Applying the DL approach to the PHEME dataset, the *N*-grams-based model surpassed LSTM by 48% in *F*-measure and 2% in accuracy.

This emphasizes the effectiveness of relying primarily on text features for optimal performance. The advantage of *N*-gram features lies in their ease of extraction from textual content. In addition to eliminating the need for extensive data corpora during the training process, as required by handcrafted features such as word embeddings, the exceptional performance of *N*-gram features stems from their ability to discern words and phrases effectively, enabling the model to grasp contextual information surrounding each word. This paper aims to fill the existing research gaps by studying the effect of the DA back-translation technique on the fake/misinformation detection task. The study presents various representations for tweet content to identify the most effective one. Moreover, it leverages the high performance of the Arab-BERT model to enhance detection performance compared to traditional machine-learning algorithms.

3 The proposed framework for Arabic fake tweet detection

The introduced framework aims to classify social media text as either fake or not. In this section, the components of the proposed framework are explained in detail, as shown in Fig. 1. A two-stage methodology was employed for this investigation. The first stage aims to comprehensively examine and evaluate various representations of the utilized datasets to identify the optimal approach. Specifically, the effectiveness of *N*-grams, content-based features, and source-based features is investigated. Subsequently, these different representations undergo meticulous evaluation using machine-learning (ML) classifiers. This stage is considered as a baseline. The second stage of the proposed methodology focuses on assessing the impact of integrating transfer learning and data augmentation (DA) techniques to increase the accuracy of the detection of fake tweets. The overall architecture encompasses three key components: the datasets, the data preparation stage, and subsequent processes, including *N*-gram extraction, hyperparameter optimization for ML algorithms, and finally, DA and classification modules.

3.1 Research datasets

The experiments in this work utilize three different datasets. The proposed approach is applied to the news dataset

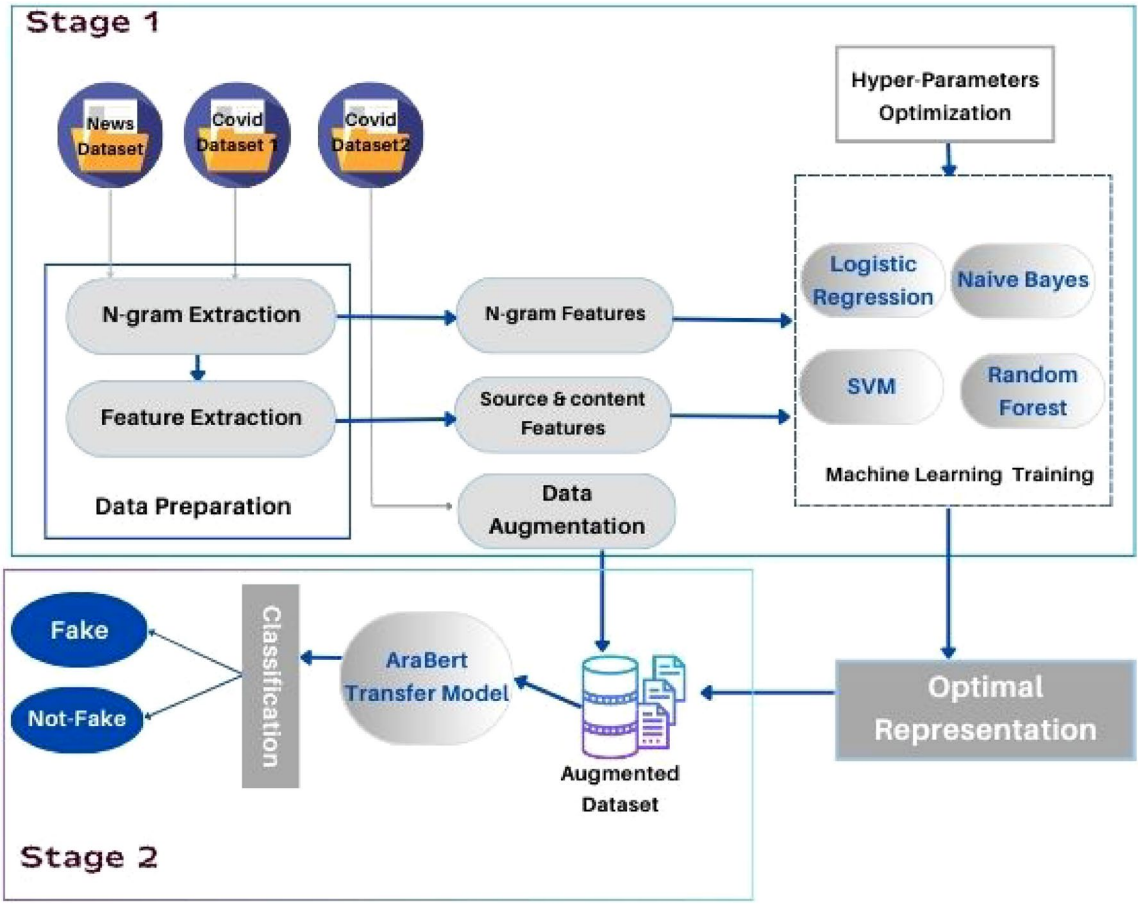

Fig. 1 The proposed framework for fake Arabic tweets detection

to make a comparison with traditional approaches in the literature.

1. News Dataset: The conducted experiments are performed using the dataset from Al Zaatari et al. (2016). This dataset comprises 1862 tweets related to the humiliation crisis in Syria. The number of fake samples is 1051, and non-fake samples are 810.
2. COVID-19 Dataset 1: The same dataset used in Alsudias and Rayson (2020) is employed in this work, consisting of Arabic tweets related to COVID-19. The dataset comprises 1,048,575 tweets, but only 2000 tweets are labeled as fake or not fake. For our experiments, we use only these 2000 labeled tweets. Labels are represented as 1 for false information, − 1 for correct information and 0 for unrelated content.
3. COVID-19 Dataset 2: For COVID-19 experiments, this work utilizes the dataset from (Mahlous and Al-Laith 2021). The dataset includes 2500 tweets about COVID-19, manually annotated into fake or non-fake classes by three annotators. After removing duplications, the annotated dataset contains 1537 tweets, with 835 labeled as fake and 702 as non-fake. The COVID dataset uses two class labels for the tweets: 0 for false information and 1 for correct information.

3.2 Data preparation and *N*-grams extraction

Data preparation is applied to Arabic tweets before *N*-gram feature extraction. This process involves removing Arabic stop words, tweet IDs, hyperlinks, emoticons, and non-Arabic words in order to reduce the size of features and improve the overall performance of the classification.

Machine-learning algorithms cannot use text data in its raw format as it is not readable for them. It must be transformed into a numerical representation or a set of vectors with the same length. These vectors are produced through a vectorization process that aims to reduce textual data into their lower dimensional space. This process is performed in two steps: the strings are tokenized to produce their corresponding tokens, and then these tokens are assigned weights to represent the importance of each token. After extracting *N*-gram features, we apply a term frequency-inverted document frequency (TF-IDF) weighting scheme,

Table 1 The optimal hyperparameters for each classifier

ML Algorithm	The tuned parameters
SVM	Kernel= linear, degree= three, gamma= one, C= one
RF	Bootstrap= False, n_estimators= five hundred
LR	C= five, Solver=newton-cg

while the *N*-gram model depends on sequences of words or characters called *N*-grams, where "*N*" represents the number of elements in a sequence. This work uses word sequences for *N*-gram models, where $n = 1$ for producing uni-grams (the bag-of-words BOW) and $n = 2$ for producing bi-grams." For example, الهواء عبر ينتقل الجديد كورونا فيروس , for this tweet the word uni-grams are الهواء , عبر , ينتقل , الجديد , كورونا , فيروس and bigrams are: الهواء عبر , عبر ينتقل , ينتقل الجديد , الجديد كورونا , كورونا فيروس and so on.

Generating various sets of *N*-gram frequencies aims to determine the optimal value of *N*. This study examines one-gram, bi-gram (2 g), and tri-gram features to evaluate their impact on the accuracy of different classification algorithms. The test tweets are pre-processed to extract the *N*-gram features, which are then fed into the generated model to predict whether new input tweets are fake or not fake.

3.3 Feature extraction

This module aims to extract source and content-based features to evaluate the proposed model using them and then compare its performance with *N*-gram features. The unigram model, also known as the bag-of-words model, is a type of language model that represents text by considering individual words as independent components. However, the unigram model does not account for the order and context of words. During feature extraction, the tweet's content is

Table 2 The statistics of augmented dataset

No	Fake tweets	Non-fake tweets
Tweets	1284	1081
Words	30,637	30,075
Unique words	7798	8706

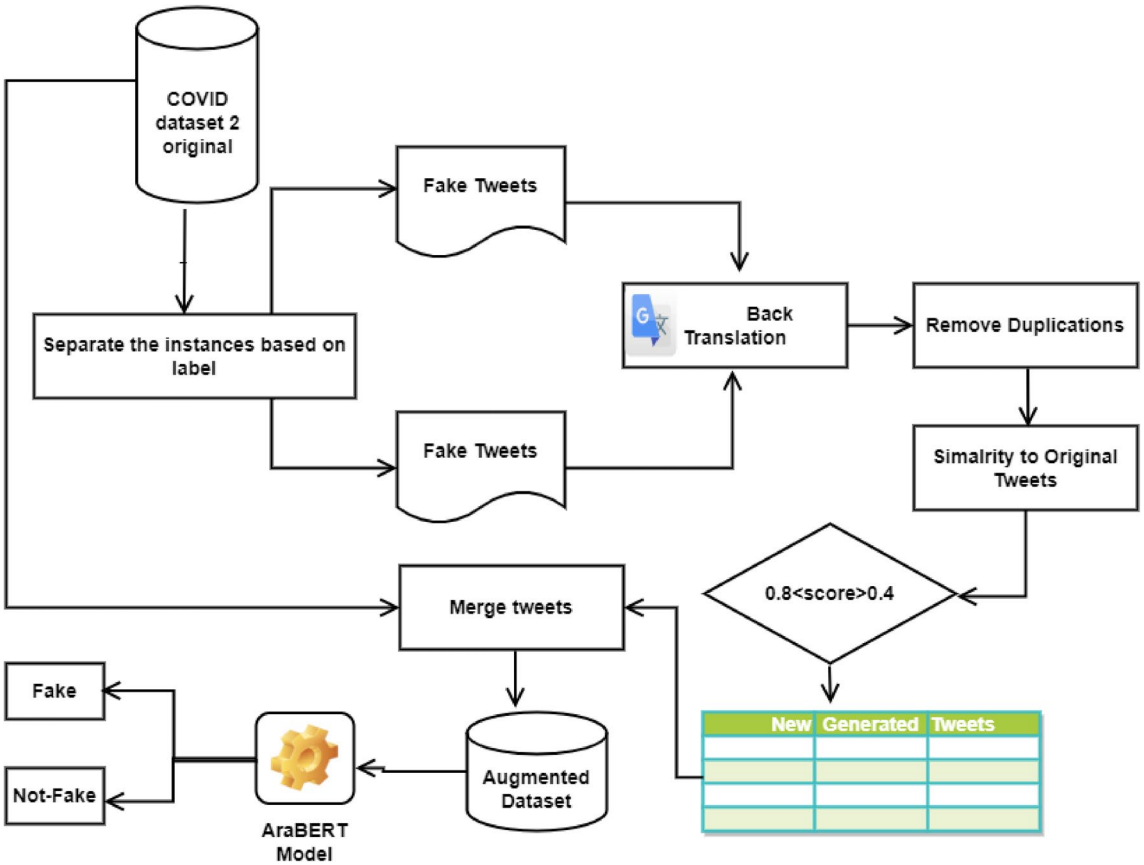

Fig. 2 WorkFlow of DA and classification with AraBERT

Table 3 Samples of generated new tweets

Original tweet	New generated tweet
فيروس كورونا في ايران الاشخاص المصابون ينهارون في الشارع هذه هي الدولة التي تعامل اوباما معها كشريك جديد مثير للولايات المتحدة في تحقيق الاستقرار في الشرق الاوسط	فيروس كورونا في إيران ، المصاب الذي ينهار في الشارع.هذا هو البلد الذي يتعين على أوباما أن يفعله كشريك جديد مثير للولايات المتحدة في الاستقرار في الشرق الأوسط
هل يستعيد بيل غيتس لقب اغنى رجل في العالم بفضل لقاح كورونا المستجد يسيطر بيل غيتس على مفاصل القطاع الصحي العالمي من خلال شبكة هائلة من الشركات والمنظمات التي يمولها بما فيها منظمة الصحة العالمية	قام بيل غيتس بإعادة لقب أغنى شخص في العالم من خلال لقاح هالة كهربائي جديد. السيطرة على مفاصل وزارة الصحة العالمية

processed to produce a feature vector for each tweet, where each element represents the occurrence or frequency of a specific word from the vocabulary. Models based on unigrams are simple and computationally efficient, but they have limitations as they may not capture the semantic information of the text. To overcome this limitation, the devised model has been extended using bi-grams.

3.4 Fine-tuning baseline machine-learning models

A set of supervised learning algorithms is introduced, providing a brief explanation of how each classifier can be applied in this context:

1. Logistic Regression (LR): In linear classifiers, logistic regression models the relationship between input features and binary outcomes (fake or not) (Maulud and Abdulazeez 2020). Arabic fake tweet detection can be achieved by training logistic regression (LR) on a labeled dataset, where features are extracted from tweets (e.g., TF-IDF vectors), and labels indicate whether a tweet is fake or not. LR learns the coefficients based on the feature representations, enabling it to predict the class of new, unseen tweets.
2. Random Forest (RF): A random forest consists of multiple decision trees combined into an ensemble classifier (Probst et al. 2019). The final prediction is determined by training each decision tree using a random subset of the samples and features, and then accumulating the predictions from each tree. With the trees learning to recognize patterns in the data based on various combinations of features, the ensemble nature of random forests (RF) brings a key benefit of reducing overfitting and improving generalization.
3. Support Vector Machine (SVM): This classifier aims to find an optimal hyperplane that separates the data points by a maximum margin that distinguishes different target classes (Huang et al. 2018). SVM can be trained to detect fake Arabic tweets using feature vectors derived from the tweets. To distinguish between fake and non-fake tweets in the feature space, SVM seeks to find the most optimal hyperplane. It can capture complex decision boundaries and effectively classify new tweets as fake or authentic by transforming the data into a higher-dimensional space.
4. Naïve Bayes (NB): In probabilistic classification, Naive Bayes uses Bayes' theorem to develop a classifier. The model assumes that the features are conditionally independent based on the class label (Reddy et al. 2022). NB can be trained on labeled data for detecting Arabic fake tweets by estimating the conditional probabilities of each feature given the class (fake or non-fake). Using the observed features of new tweets, NB calculates the posterior probabilities for each class and selects the most likely class. Naïve Bayes has been demonstrated to be computationally efficient and capable of handling large feature spaces, making it a suitable method for text analysis.

This type of classifier is trained on labeled data, where features extracted from Arabic tweets are combined with 1 and 2 g features using the TF-IDF weighting scheme. After training, the models are employed to classify new, unseen tweets and determine their authenticity based on the learned patterns and decision boundaries. Ultimately, the best-performing classifier is utilized for tweet classification.

In addition, hyperparameter tuning is performed for each classifier. Hyperparameters, which are not directly learned within the estimators but impact the model's performance, are adjusted to address overfitting and underfitting issues. The grid search module from the scikit-learn library optimizes the classifiers' performance. It exhaustively tries all combinations of hyperparameters for ML algorithms to select the optimal values. The module builds models for each combination and selects the best one based on evaluation

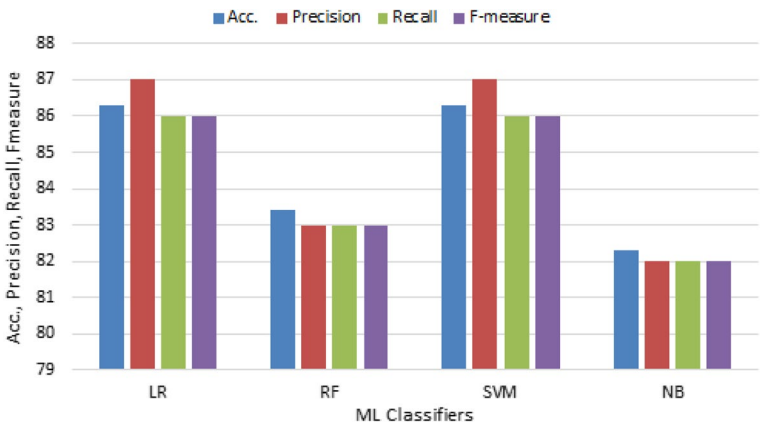

Fig. 3 Accuracy of the combination of uni-gram and bi-gram

metrics. Table 1 presents the tuned optimal parameters for the classifiers used, while default values were utilized for the other hyperparameters.

3.5 Integration of data augmentation and transfer learning (AraBERT) for fake tweets detection enhancement

This section presents the combination of the DA back-translation technique, natural language processing, and transfer learning with the AraBERT model, which has been employed to enhance the performance of Arabic fake tweet detection.

The steps of the augmentation process are illustrated in Fig. 2. The original tweets are divided into two categories: fake and non-fake tweets. Subsequently, a back-translation task is applied to both sets of tweets to generate the corresponding back-translated versions. This task employs two pivot languages, English and French. The Arabic tweets are translated into English and French and then back-translated into Arabic.

After removing duplicated tweets, the cosine similarity is calculated between the original and back-translated tweets. If the similarity score falls between the 0.4 and 0.8 thresholds, the back-translated tweet is added to the new augmented dataset. The newly generated tweets are then merged with the original tweets to create the augmented dataset. Table 2 presents the statistics of the augmented dataset, while Table 3 provides three examples of the newly generated tweets.

The final step involves classifying tweets as either fake or not fake using the pre-trained Arabic-Bert model. Arabic-Bert is selected for its high performance in detecting Arabic fake tweets (Refai et al. 2022). The commonly used Arabic-BERT pre-trained models include mini-BERT, medium-BERT, base-BERT, and large-BERT (Devlin et al. 2018). *Mini-BERT*: The mini-BERT model was designed in order to have a smaller memory consumption and a smaller size than the original BERT model (Almaliki et al. 2023). This can be achieved by reducing either the number of layers or the hidden size of the transformer model. In addition to being more computationally efficient, mini-BERT requires fewer computational resources during training and inference due to its downsized architecture. Even though it may not perform as well as larger BERT models, it can still capture contextual information and semantic relationships within a text in Arabic.

Medium-BERT: An alternative to Arabic-BERT, medium-BERT strikes a balance between size and performance. The mini-BERT model strives to achieve a reasonable compromise between the computational requirements of the full BERT model and the smaller mini-BERT model (Devlin et al. 2018; Chouikhi et al. 2021). Medium-BERT models contain fewer layers than full BERT models, and thus, they are more resource-efficient for retaining high performance for NLP tasks. The mini-BERT and medium-BERT models are tailored versions of the original BERT model specifically designed for processing Arabic texts (Chouikhi et al. 2021). Due to their lightweight and efficient nature, they are suitable for the main target of this study.

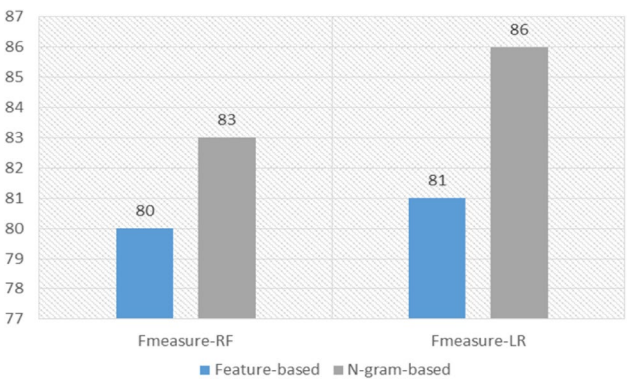

Fig. 4 Feature-based versus *N*-gram-based *F*-measure

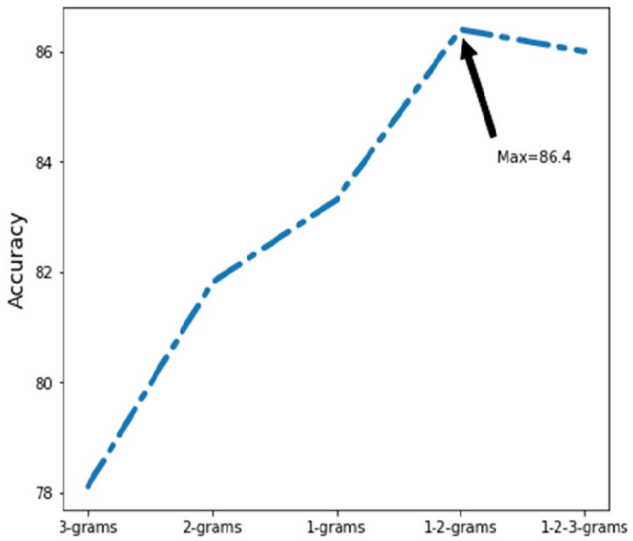

Fig. 5 Accuracy of uni-gram, bi-gram, and tri-gram

Table 5 Results of using (1 + 2 g) for Covid Dataset 1

	Accuracy	Precision	Recall	*F*-Measure
LR	86.4	86.3	80.7	83.0
RF	82.1	81	75	77
SVM	86.4	86	79.7	82
NB	81.5	81	76.7	78.3

Table 4 Feature-related versus the proposed *N*-gram model

Model	Accuracy	Precision	Recall	*F*-measure
LR for Feature	76.8	77.9	84.9	81
LR for *N*-gram	86.3	87	86	86
RF for Feature	76	78	82	80
RF for *N*-gram	83.4	83	83	83

4 Evaluation metrics

This work utilized metrics such as accuracy, precision, recall, *F*-measure, loss, ROC AUC curve, and confusion matrix to evaluate the performance of the proposed framework. The ROC AUC score (AUC) specifically measures the model's ability to distinguish between positive and negative classes. The performance evaluation metrics for classifiers are defined in the following equations:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}. \quad (1)$$

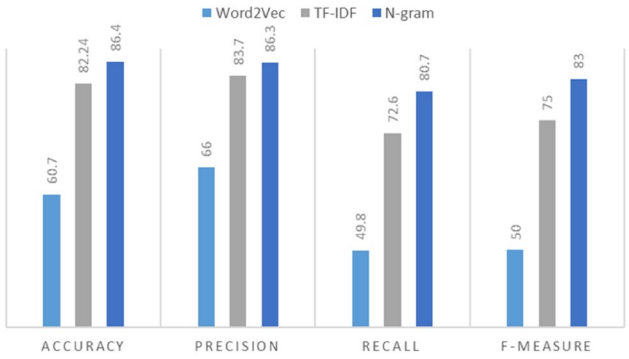

Fig. 6 LR results

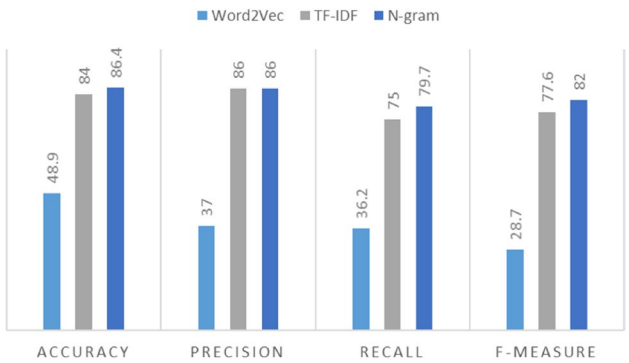

Fig. 7 SVM results

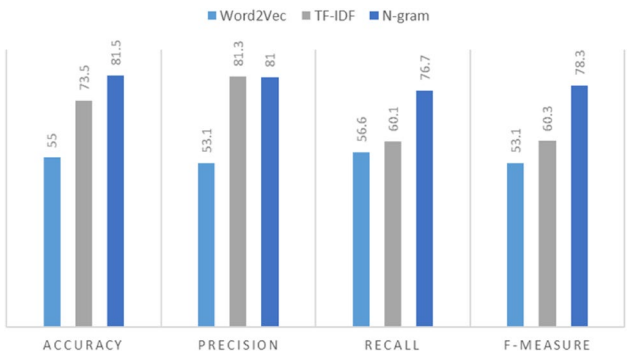

Fig. 8 NB results

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (2)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3)$$

Table 6 The statistics of original dataset

No	Fake tweets	Non-fake tweets
Tweets	835	702
Words	20,967	20,224
Unique Words	6246	7216

Table 7 The statistics of augmented dataset

No	Fake tweets	Non-fake tweets
Tweets	1284	1081
Words	30,637	30,075
Unique Words	7798	8706

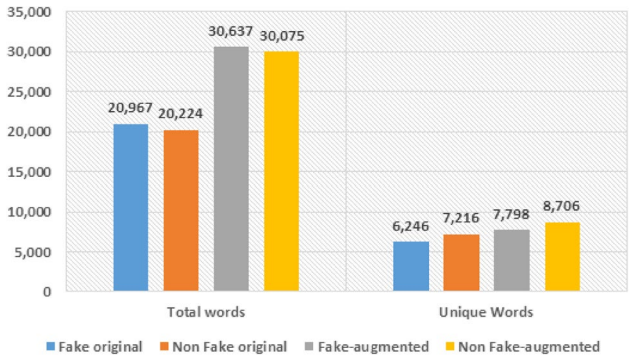

Fig. 9 Diversity of original and augmented datasets

$$F\text{-measure} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \quad (4)$$

TP represents true positives. TN represents the true negative number. FP represents false positives. FN represents false negatives.

5 Experimental results and disscusion

This study encompassed two experiments. The initial experiment, detailed in Sect. 5.1, explored various representations of datasets, encompassing *N*-gram, content, and source features. These representations went through assessment via machine-learning classifiers to ascertain the optimal approach. Subsequently, the second experiment (Sect. 5.2) delved into examining the influence of integrating transfer learning and data augmentation techniques to augment the detection of Arabic fake tweets.

To enhance the precision of the learning process, we employed fivefold cross-validation. This method involves partitioning the dataset into fivefolds, with the model utilizing four for training and the remaining one for performance evaluation.

5.1 Results of various dataset representations

In this subsection, two experiments will be discussed. In Experiment A, we highlight the findings of the *N*-gram model applied to the news dataset, as mentioned in Sect. 1. In Experiment B, the proposed *N*-gram-based model is evaluated using a dataset comprising COVID-19 Arabic tweets, as detailed in Sect. 2.

5.1.1 Experiment A: results using the news dataset

The results of the proposed model for both uni-grams and bi-grams representations are displayed in Fig. 3. Overall, the LR and SVM algorithms demonstrate the best performance with an accuracy of 86.3%, precision of 87%, recall of 86%, and *F*-measure of 86%. On the other hand, the NB algorithm displays the lowest results with an accuracy of 82.3%. Figure 3 displays the accuracy of various classifiers (NB, LR, RF, SVM), with the LR classifier achieving the highest accuracy of 86.3%. This notable performance is attributed to the inclusion of bi-grams alongside uni-gram features,

Fig. 10 Word cloud for original dataset

Fig. 11 Word cloud for augmented dataset

Table 8 Results of (Original dataset)

Classifier	Accuracy	Precision	Recall	*F*-measure
NB	66.52	66.23	68.50	67.10
LR	75.13	75.32	76.80	75.60
SVM	72.84	72.73	73.20	72.94
RF	76.85	76.62	76.95	77.13
Mini-BERT	81.46	81.85	80.92	81.08
Meduim-BERT	83.53	83.21	83.15	83.00

Table 9 Results of augmented dataset

Classifier	Accuracy	Precision	Recall	*F*-measure
NB	76.15	76.37	76.50	76.40
LR	78.50	78.90	78.80	78.85
SVM	79.90	79.75	80.05	80.10
RF	81.30	81.02	82.25	81.40
Mini-BERT	93.35	93.63	94.12	93.20
Meduim-BERT	95.46	95.52	95.30	94.82

enabling effective discrimination of words and phrases. By incorporating bi-grams, the model gains contextual information surrounding each word. Furthermore, the combination of uni-grams and bi-grams contributes to the overall improvement in detection performance. A recent study by Jardaneh et al. (2019) introduced a model utilizing content and user-related features. However, the model presented in Jardaneh et al. (2019) strictly relies on textual features, such as word frequency, which neglects the relationships between words. To compare our *N*-gram-based model with state-of-the-art content-related and source-related models presented in Jardaneh et al. (2019), both models are applied to the dataset used in Al Zaatari et al. (2016). The results of these two models are shown in Table 4. From the results, we can conclude that our proposed *N*-gram model is more effective than the feature-related model. This suggests that incorporating *N*-grams gives our model the ability to capture meaningful word relationships and contextual information, contributing to its superior performance in detecting fake Arabic tweets.

Furthermore, LR and RF algorithms employed by Jardaneh et al. (2019) are utilized with a combination of uni-grams and bi-grams weights. From Fig. 4 and Table 4, it is evident that the proposed *N*-gram model outperforms the feature-based model of Jardaneh et al. (2019), particularly when using LR and RF classifiers. Specifically, the *N*-gram-based LR achieves the highest performance, surpassing the feature-based LR by 9.3% in accuracy, 10% in precision, 1% in recall, and 5% in *F*-measure. The superiority of the *N*-gram model over the feature-based model of Jardaneh et al. (2019) is attributed to the utilization of bi-grams combined with uni-grams, representing the context of words as phrases. Additionally, Fig. 4 suggests that hyperparameter tuning plays a crucial role in improving classifier performance and mitigating the issues of overfitting and underfitting, ultimately leading to higher *F*-measure scores.

5.1.2 Experiment B: results using Covid-19 dataset

The approach outlined in Alsudias and Rayson (2020) serves as a baseline, ensuring a fair comparison. We employed the same machine-learning (ML) classifiers utilized in Alsudias and Rayson (2020), namely SVM, NB, and LR, and applied the identical TF-IDF weighting scheme.

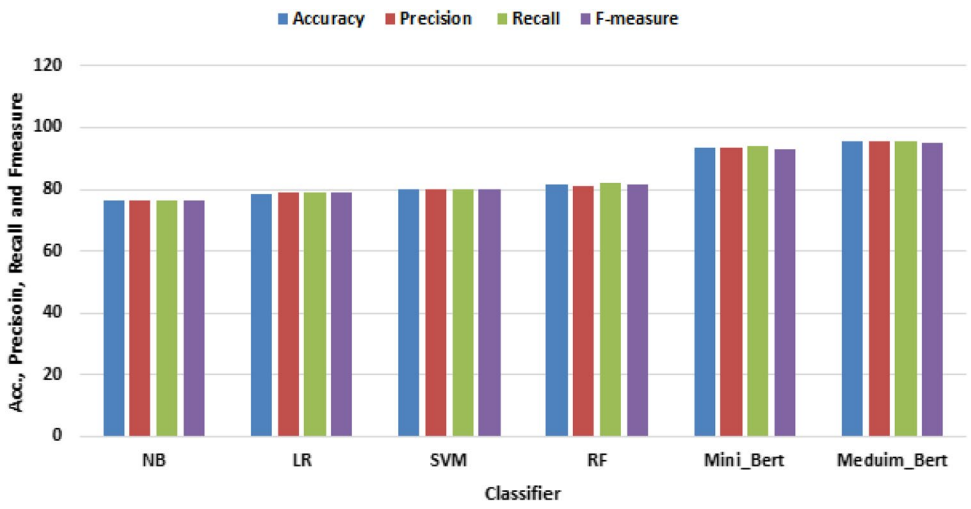

Fig. 12 AraBERT improvements to baselines ML classifiers

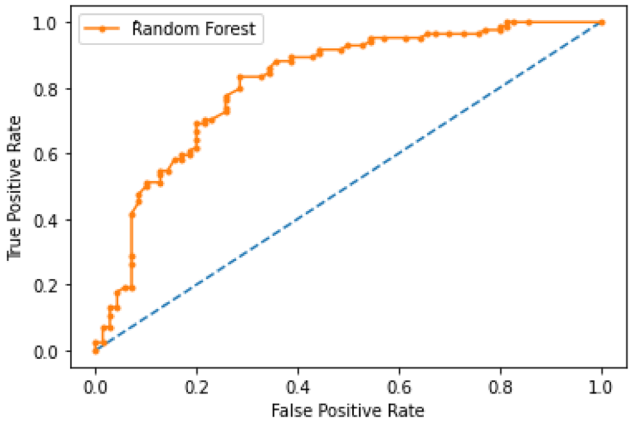

Fig. 13 Random Forest ROC curve analysis for Original dataset

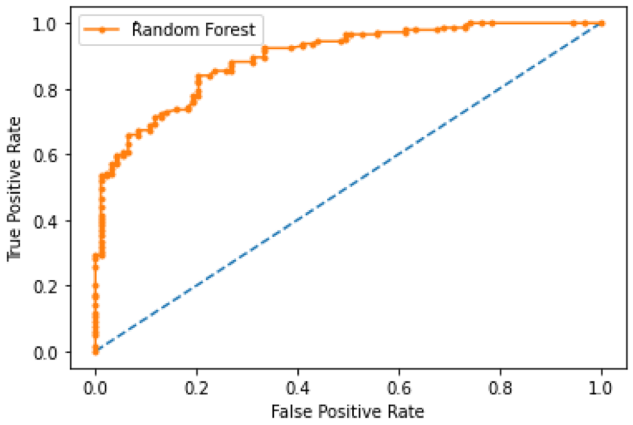

Fig. 14 Random Forest ROC curve analysis for the Augmented dataset

Furthermore, a word-based *N*-gram analysis was conducted to determine the optimal value of *N* for detecting misinformation in the target COVID dataset. Exploring the effect of *N*-gram length on the classifier's performance, N was set to one, two, and three. The results, illustrated in Fig. 5, reveal promising outcomes for the uni-gram, bi-gram, and tri-gram models. Combining uni-grams and bi-grams, however, achieved the highest performance, attaining the best accuracy of 86.4%, as demonstrated in Fig. 5.

For the SVM, NB, and LR algorithms, the results of the *N*-gram model are summarized in Table 5. Notably, on the Covid-19 dataset 1, LR-N-gram consistently demonstrates the best performance.

Additionally, the results of the *N*-gram model are compared with those presented in Alsudias and Rayson (2020), where TF-IDF-based frequency features and Word2Vec-based embedding features were employed. LR, RF, and SVM were selected as the classifiers. A detailed comparison of the proposed *N*-gram classifier against TF-IDF and Word2Vec is illustrated in Figs. 6, 7, and 8 for the LR, RF, and SVM classifiers.

Combining LR with *N*-grams leads to superior performance compared to LR using TF-IDF and Word2Vec alone. Figure 6 illustrates that LR *N*-gram-based achieves an accuracy of 86.4%, recall of 80.7%, and an *F*-measure of 83%. The results from Figs. 6, 7, and 8 collectively demonstrate that utilizing *N*-gram features for all three classifiers is more effective than using either Word2Vec or TF-IDF features alone. Moreover, the process of creating *N*-gram features is simpler compared to generating Word2Vec, which requires additional resources and pre-training of word embeddings.

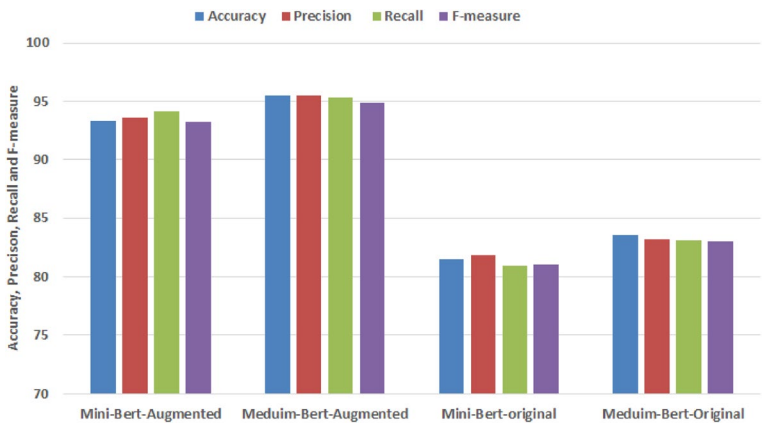

Fig. 15 Comparison of results between the original and augmented datasets

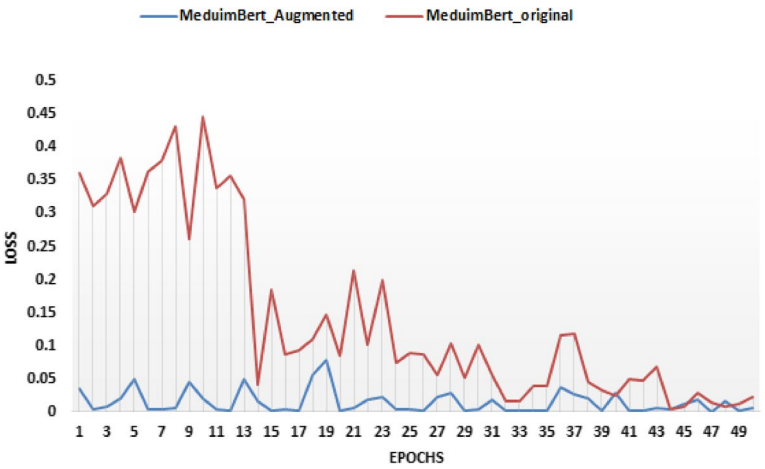

Fig. 16 AraBERT loss for original dataset and augmented datasets

5.2 Data augmentation results

The COVID-19 dataset 2 (as described in Sect. 3) consists of 835 instances of fake classes and 702 instances of non-fake classes. Tables 6 and 7 illustrate the statistics of the original dataset and the augmented dataset. The data augmentation (DA) process results in an increase in the amount of training data for both fake and non-fake classes. Since the back-translation DA technique provides diversity and novelty to the generated data instances, the unique words in the augmented dataset have increased by 1552 for the fake class and 1490 for the non-fake class compared to the original dataset. Additionally, the number of new instances has increased by 449 for the fake class and 379 for the non-fake class. Figure 9 illustrates that the diversity is highest for the non-fake class in the augmented dataset. Moreover, Figs. 10 and 11 show the frequencies of words for original and augmented datasets. It is clear that some words frequencies have changed and new words appeared after augmentation process such as (مجففات) word.

To investigate the effect of the data augmentation task, the results of both the original and augmented datasets are provided in Tables 8 and 9. It is clear that the RF classifier outperforms NB, LR, and SVM classifiers, achieving an *F*-measure of 76.62%. When using the augmented dataset, the *F*-measure of RF increases by 4.4% compared to the original dataset.

According to the results for the AraBERT model displayed in Tables 8 and 9, the findings demonstrate that transfer learning models are highly accurate at spotting auto-generated texts. Additionally, all baseline models were significantly outperformed by the AraBERT model. Our proposed AraBERT model outperforms the baselines by +17% for LR, +19.3% for NB, +14.2% for RF, and +15.6% for SVM. Figure 12 compares the outcomes to better illustrate the improvements.

Figures 15 and 16 illustrate the results of mini-BERT and medium-BERT for both the original and augmented datasets. Classification performance using the augmented dataset is better than the original dataset. Moreover, the performance of mini-BERT and medium-BERT surpasses ML models. In particular, medium-BERT achieves the highest performance among all models for both the augmented and original datasets. Furthermore, the AUC of all classifiers is increased by

around 7 to 15%. Figures 13 and 14 show the ROC curve analysis of the random forest classifier for the original and augmented datasets. In future work, different values for the threshold can be checked, and their effect on the size of the augmented dataset and the classification performance can be studied. The dataset augmentation stage can be implemented using other transformers, such as the GPT2 model, which considers label preserving during the augmentation process (Figs. 15, 16).

6 Conclusions and future work

In this study, Arabic fake tweet detection is enhanced through the integration of transfer learning and domain adaptation techniques. The evaluation of *N*-grams as model features reveals that a combination of two and three grams outperforms content and source-based models, demonstrating improved machine-learning performance. Logistic regression (LR) with *N*-grams achieves notable accuracy (86.3%) in fake news detection, surpassing Word2Vec and TF-IDF-based models. The study also explores the impact of back-translation augmentation, revealing superior performance on augmented datasets when employing mini-BERT and medium-BERT classifiers. Medium-BERT consistently outperforms traditional ML models, enhancing the area under the curve (AUC) by 7 to 15%. The integration of back-translation with pre-trained models proves pivotal in improving the classification of Arabic tweets, emphasizing a contextual approach over manual feature crafting. As a forward-looking suggestion, future work is recommended to explore diverse domain adaptation approaches and assess the potential benefits of incorporating generative methods for further advancements in Arabic fake tweet detection.

Acknowledgements We thank Minia University and Stdf for the financial support for publishing this work.

Author contributions EM: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Writing the paper. WNI: Conceived and designed the experiments; Analyzed and interpreted the data; Writing the paper. EY: Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper. OA: Analyzed and interpreted the data; analysis tools or data; Wrote the paper. All authors contributed to the manuscript and reviewed it.

Funding Open access funding provided by The Science, Technology & Innovation Funding Authority (STDF) in cooperation with The Egyptian Knowledge Bank (EKB).

Availability of data and materials Data is available upon reasonable request from the corresponding author.

Declarations

Conflict of interest The authors declare no conflict of interests

References

Al-Khalifa H, Abuzayed A (2021) Sarcasm and sentiment detection in Arabic tweets using Bert-based models and data augmentation. In: Proceedings of the sixth Arabic natural language processing workshop

Ajao O, Bhowmik D, Zargari S (2018) Fake news identification on twitter with hybrid CNN and RNN models. In: Proceedings of the 9th international conference on social media and society, pp 226–230

Al-Dhabyani W, Gomaa M, Khaled H, Aly F (2019) Deep learning approaches for data augmentation and classification of breast masses using ultrasound images. Int J Adv Comput Sci Appl 10:1–11

Al Zaatari A, El Ballouli R, ELbassouni S, El-Hajj W, Hajj H, Shaban K, Habash N, Yahya E (2016) Arabic corpora for credibility analysis. In: Proceedings of the tenth international conference on language resources and evaluation (LREC'16), pp 4396–4401

Albalawi RM, Jamal AT, Khadidos AO, Alhothali AM (2023) Multimodal Arabic rumors detection. IEEE. Access 11:9716–9730

Elkorany A, Ahmed C, Alkadri AM (2022) Enhancing detection of Arabic social spam using data augmentation and machine learning. Appl Sci 12(22):11388

Almaliki M, Abdulqader AM, Ibrahim G, El-Sayed A (2023) ABMM: Arabic Bert-mini model for hate-speech detection on social media. Electronics 12:1048

Alsudias L, Rayson P (2020) COVID-19 and Arabic Twitter: How can Arab world governments and public health organizations learn from social media? In: Association for computational linguistics

Asta RS, Setiawan EB (2023) Fake news (HOAX) detection on social media using convolutional neural network (CNN) and recurrent neural network (RNN) methods. In: 2023 11th International conference on information and communication technology (ICoICT), pp 511–516. IEEE

Azad R (2023) A novel taxonomy for Arabic fake news datasets. Int J Comput Digital Syst 14(1):1–1

Bayer M, Kaufhold M-A, Reuter C (2022) A survey on data augmentation for text classification. ACM Comput Surv 55:1–39

Bayer M, Kaufhold MA, Buchhold B, Keller M, Dallmeyer J, Reuter C (2023) Data augmentation in natural language processing: a novel text generation approach for long and short text classifiers. Int J Mach Learn Cybern 14(1):135–150

Capuano N, Fenza G, Loia V, Nota FD (2023) Content based fake news detection with machine and deep learning: a systematic review. Neurocomputing

Chouikhi H, Chniter H, Jarray F (2021) Arabic sentiment analysis using Bert model. In: 13th International conference advances in computational collective intelligence, pp 621–632

Cuesta Á, Barrero DF, R-Moreno MD (2013) A descriptive analysis of twitter activity in Spanish around Boston terror attacks. In:

International conference on computational collective intelligence, pp 631–640, Springer: New York
Devlin J, Chang M-W, Lee K, Toutanova K (2018) Bert: pre-training of deep bidirectional transformers for language understanding. *arXiv preprint*arXiv:1810.04805
El Ballouli R, El-Hajj W, Ghandour A, Elbassuoni S, Hajj H, Shaban K (2017) Cat: credibility analysis of Arabic content on twitter. In: Proceedings of the third Arabic natural language processing workshop, pp 62–71
Gaanoun B-IK (2020) Arabic dialect identification: an Arabic-Bert model with data augmentation and ensembling strategy. In: Proceedings of the fifth Arabic natural language processing workshop
Gupta S, Verma B, Gupta P, Goel L, Arun KY, Yadav D (2023) Identification of fake news using deep neural network-based hybrid model. SN Comput Sci 4(5):679
Hassan N, Gomaa W, Khoriba G, Haggag M (2020) Credibility detection in Twitter using word *N*-gram analysis and supervised machine learning techniques. Int J Intel Eng Syst 13:291–300
Hassan NY, Gomaa WH, Khoriba GA, Haggag MH (2018) Supervised learning approach for Twitter credibility detection. In: 2018 13th International conference on computer engineering and systems (ICCES), pp 196–201. IEEE
Hua J, Cui X, Li X, Tang K, Zhu P (2023) Multimodal fake news detection through data augmentation-based contrastive learning. Appl Soft Comput 136:110125
Huang S, Cai N, Pacheco PP, Narrandes S, Wang Y, Xu W (2018) Applications of support vector machine (SVM) learning in cancer genomics. Cancer Genom Proteom 15:41–51
Jardaneh G, Abdelhaq H, Buzz M, Johnson D (2019) Classifying Arabic tweets based on credibility using content and user features. In: 2019 IEEE Jordan international joint conference on electrical engineering and information technology (JEEIT), pp 596–601. IEEE
Jiang W, Ling L, Zhang D, Lin R, Zeng L (2023) A time series forecasting model selection framework using CNN and data augmentation for small sample data. In: Neural processing letters, pp 1–28
Jin Z, Cao J, Jiang Y-G, Zhang Y (2014) News credibility evaluation on microblog with a hierarchical propagation model. In: 2014 IEEE International conference on data mining, pp 230–239. IEEE
Kazmi M, Habib S, Hayat S, Rehman L, Aziz A, Qazi SA (2023) Unihach: unicode and hash function supported with counting and frequency recurrence of Arabic characters for Quranic text watermarking. Arab J Sci Eng 1–17
Kumar V, Choudhary A, Cho E (2020) Data augmentation using pre-trained transformer models. *arXiv preprint*arXiv:2003.02245
Li B, Hou Y, Che Wanxiang (2022) Data augmentation approaches in natural language processing: a survey. Ai Open 3:71–90
Li G, Wang H, Ding Y, Zhou K, Yan X (2023) Data augmentation for aspect-based sentiment analysis. Int J Mach Learn Cybern 14(1):125–133
Lorek K, Suehiro-Wiciński J, Jankowski-Lorek M, Gupta Amit (2015) Automated credibility assessment on twitter. Comput Sci 16:157–168
Lu X, Brelsford C (2014) Network structure and community evolution on twitter: human behavior change in response to the 2011 Japanese Earthquake and Tsunami. Sci Rep 4:1–11
Mahlous AR, Al-Laith A (2021) Fake news detection in Arabic tweets during the Covid-19 pandemic. Int J Adv Comput Sci Appl 12(6):778–788
Maulud D, Abdulazeez AM (2020) A review on linear regression comprehensive in machine learning. J Appl Sci Technol Trends 1:140–147
Mohamed E, Elmougy S, Aref M (2019) Toward multi-lingual information retrieval system based on internet linguistic diversity measurement. Ain Shams Eng J 10:489–497
Mohamed E, Elmougy S, Ali-Sadek IO, Aref M (2019b) Semantic relatedness based query translation disambiguation approach for cross-language web search. Int J Adv Sci Technol
Mourad A, Srour A, Harmanai H, Jenainati C, Arafeh M (2020) Critical impact of social networks Infodemic on defeating coronavirus Covid-19 pandemic: Twitter-based study and research directions. IEEE Trans Network Serv Manag 17:2145–2155
Pellicer LFAO, Ferreira TM, Costa AHR (2023) Data augmentation techniques in natural language processing. Appl Soft Comput 132:109803
Probst P, Wright MN, Boulesteix A-L (2019) Hyperparameters and tuning strategies for random forest. Wiley interdisciplinary reviews: data mining and knowledge discovery 9:e1301
Ravikumar S, Balakrishnan R, Kambhampati S (2012) Ranking tweets considering trust and relevance. In: Proceedings of the ninth international workshop on information integration on the web, pp 1–4
Reddy EMK, Gurrala A, Hasitha VB, Kumar KVR (2022) Introduction to Naive Bayes and a review on its subtypes with applications. In: Bayesian reasoning and Gaussian processes for machine learning applications, pp 1–14
Refai D, Abo-Soud S, Abdel-Rahman M (2022) Data augmentation using transformers and similarity measures for improving Arabic text classification. *arXiv preprint*arXiv:2212.13939
Sabbeh SF, Baatwah SY (2018) Arabic news credibility on twitter: an enhanced model using hybrid features. J Theor Appl Inform Technol 96:2327–2338
Sabty C, Omar I, Wasfalla F, Islam M, Abdennadher S (2021) Data augmentation techniques on Arabic data for named entity recognition. Proc Comput Sci 89:292–299. https://doi.org/10.1016/j.procs.2021.05.092
Singh MK, Ahmed J, Alam MA, Raghuvanshi KK, Kumar S (2023) A comprehensive review on automatic detection of fake news on social media. Multimed Tools Appl 1–34
Wang H, Duentsch I, Guo G, Khan SA (2023) Special issue on small data analytics. Int J Mach Learn Cybern 14(1):1–2
Yadav A, Vishwakarma DK (2023) MRT-net: Auto-adaptive weighting of manipulation residuals and texture clues for face manipulation detection. Exp Syst Appl 232:120898
Yadav AK, Suraj K, Dipesh K, Lalit K, Kapil K, Maurya SK, Mohit K, Divakar Y (2023) Fake news detection using hybrid deep learning method. SN Comput Sci 4(6):845
Zubiaga A, Liakata M, Procter R (2017) Exploiting context for rumour detection in social media. In: International conference on social informatics, pp 109–123, Springer: New York