Engineering Applications of Artificial Intelligence 126 (2023) 106999

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Sentiment analysis of imbalanced datasets using BERT and ensemble stacking for deep learning

Nassera Habbat [a,*], Hicham Nouri [b], Houda Anoun [a], Larbi Hassouni [a]

[a] RITM Laboratory, CED ENSEM Ecole Superieure de Technologie Hassan II University, Casablanca, Morocco
[b] Research Laboratory on New Economy and Development (LARNED), Faculty of Legal Economic and Social Sciences AIN SEBAA, Hassan II University, Casablanca, Morocco

ARTICLE INFO

Keywords:
BERT
Deep learning
Ensemble learning
Imbalanced datasets
Sentiment analysis
SMOTE

ABSTRACT

The Internet is a crucial way to share information in both personal and professional areas. Sentiment analysis attracts great interest in marketing, research, and business today. The instability faced by imbalanced datasets on sentiment analysis is examined in this research. Balancing the datasets using techniques based on under-sampling and over-sampling is examined to achieve more efficient classification results as the effects of using BERT as word embedding and ensemble learning methods for classification. The effects of the resampling training set algorithms on different deep learning classifiers were investigated using BERT as a word embedding model and Cohen's kappa, accuracy, ROC-AUC curve, and MCC as evaluation metrics with k-fold validation on three sentiment analysis datasets containing English, Arabic, and Moroccan Arabic Dialect texts. Also, we did those performance metrics for all models when scaling the dataset for training and testing, and we calculated the memory and the execution time for each model. Finally, we analyzed the National Office of Railways of Morocco (ONCF) customers' Facebook comments in Modern Standard Arabic (MSA) and MD to determine customer satisfaction as positive, negative, and neutral comments.

1. Introduction

Different tools like forums, e-commerce websites, and social media were established with Web 2.0, allowing people to share their thoughts and reviews on various topics concerning products, events, and services. Automatically analyzing and examining this unstructured sharing data to obtain important information about users' thoughts and behaviors is crucial for several individuals and organizations, including businesses, customers, and governments. Therefore, numerous research topics, like big data and sentiment analysis, have grown increasingly active.

Sentiment analysis, often known as SA, is a natural language processing (NLP) activity that has amassed a substantial significance over the past several years in information extraction and data analysis. The primary purpose of SA is to identify the emotions expressed in text and categorize the polarities of these emotions as having either a ternary or binary orientation. Analysts increasingly turn to social media as their primary data source to investigate internet users' opinions on a specific theme, enabling them to predict and change their plans. Many machine learning algorithms for sentiment analysis have been developed, particularly sequence models that encapsulate the text's long-distance relationships. However, the sequence models could be more computationally efficient if the processing is serialized. The RNN-based modeling methods are widely used in various applications, including energy and medicine (Skrobek et al., 2020, 2022; Q. G. Xu et al., 2019).

Most standard data classification algorithms can be applied proficiently in terms of overall classification accuracy if the data in each class is equally distributed. Nevertheless, these classifiers could not perform better when classifying imbalanced data with low instances in the interest group. For most learning algorithms, imbalanced data violates the assumption of a relative equilibrium distribution, which can significantly reduce classification performance.

Regarding resources, most dialects of Dialectal Arabic are regarded as low-resource languages. There needs to be more labeled data to build NLP systems for them; consequently, more research is needed. Also, most of the work done before has been on MSA, and a few dialects, such as Egyptian and Middle east regions. Moroccan Darija belongs to the Maghrebi dialects group (Hassine et al., 2016) and is known as Morocco's Arabic colloquial. Moroccan Darija is a dialect with more than 36 million native speakers and is the primary dialect spoken in daily interactions, informal situations, and routine activities.

* Corresponding author.
E-mail address: nassera.habbat@gmail.com (N. Habbat).

https://doi.org/10.1016/j.engappai.2023.106999
Received 30 January 2023; Received in revised form 20 July 2023; Accepted 14 August 2023
Available online 30 August 2023

Morocco had 20 million social media users in 2021 (Statista The Statistics Portal, 2011), 23% more than in 2019. But, more work needs to be done on the Moroccan Arabic dialect. For that, we aim in this work to explore sentiment analysis on imbalanced datasets in three languages; English, standard Arabic, and the Morrocan Dialect using different techniques for data resampling, Bidirectional Encoder Representations from Transformers (BERT) as a contextualized work embedding, and an ensemble stacked deep learning. Briefly, the following are the main findings of this paper.

1. We used imbalanced datasets in three different languages (English, Arabic standard, and Moroccan Arabic dialect),
2. We compared over-sampling and under-sampling techniques to balance the datasets,
3. We used BERT as contextualized word embedding to extract features,
4. We classified the datasets using the stacking deep learning model, which combines three deep learning algorithms (Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), and Bidirectional LSTM (BiLSTM)) using Support Vector Machine (SVM) as a meta-classifier.

The rest of this article is the following: the second section offers an overview of Arabic sentiment analysis works. In the third section, we describe the proposed approach; then, the experiment findings are addressed in section 4. Finally, the conclusion is drawn in the last section.

2. Related works

The majority of the techniques observed for Arabic and dialect sentiment classification rely on traditional Machine Learning (Hicham et al., 2022; Hicham and Karim, 2022; Jihad et al., 2023; Omari, 2022) used Logistic Regression (LR) as an ML algorithm for Arabic SA to classify customer reviews, and the used algorithm achieved the best result using TF-IDF (term frequency and inverse document frequency). In the same context (Hadwan et al., 2022), utilized different ML algorithms to analyze the opinions of Saudi Arabian people on social media and other applications such as Google Play. They found in their experiments that K-Nearest Neighbor (KNN) was the best, with an accuracy of 78.46% compared to SVM, Naïve Bayes (NB), and Decision Tree (DT). NB, Rocchio classifier, and SVM are stacked by (Omar et al., 2013) for both Subjectivity and SA of Customers Arabic Reviews. The result demonstrated that the ensemble model proposed obtained the best results using LR as the meta-learner ensemble technique.

On the contrary, DL techniques for SA (Elnagar et al., 2020; Farha and Magdy, 2021), including recurrent neural networks (RNN) (Williams and Zipser, 1989) and convolutional neural networks (CNN) (Zhang et al., 2011), have been explored in recent years as providing the capability to offer higher adaptability and robustness by extracting features automatically. However, compared to its applications in other domains, such as chatbots, recommendation systems, remote sensing, and load monitoring, deep neural network (DNN) methods in Arabic Dialect Sentiment Analysis still need to be improved.

(Mhamed et al., 2021) used CNN to classify Arabic tweets, and they obtained an accuracy of up to 90.06% utilizing 2 CNN layers, a global average pooling layer, and a dense layer with Word2vec word embedding. However (M.Abdelgwad et al., 2021), chose to use the RNN model to classify Arabic hotel reviews, and they got the best performance using the Attention network based on BiGRU–CNN–CRF with an accuracy of 83.98% (Alali et al., 2022). presented a hierarchical attention network-based multitask learning model named MTLHAN. Their model showed the best performance with high accuracy, superior to 83%, on three Arabic datasets. Recently, an ensemble of deep learning models showed a good performance (Abu Kwaik et al., 2019); introduced a DNN that combines Bi-LSTM with CNN using word2vec embeddings for the binary and three-way classification Arabic text. The proposed combination gave the most accurate outcome with an accuracy of up to 81% and 93% for binary classification and between 66% and 77% for three-way classification. Similarly (Habbat et al., 2022), proposed a combination of CNN and GRU using XLNet as a word embedding (WE) model.

Several efforts have recently been implemented in the low-resource datasets to cover more DA variations (Boujou et al., 2021); presented a labeled dataset of 50K tweets in five Arabic dialects, including the Moroccan Dialect with 9965 tweets. The authors conducted different experiments on this dataset, including the sentiment analysis task using ML algorithms (SGD Classifier, LR, NB, and Linear SVC). They found that in terms of accuracy and f1-score, the SGD Classifier is a bit better. Concerning Moroccan Dialect (Garouani and Kharroubi, 2022a); has published an open-access dataset that analyzes sentiments for the Moroccan Dialect. The data was mainly 18000 collected tweets and manually annotated. The authors applied different experiments (Garouani and Kharroubi, 2022b) to this dataset, comparing machine learning, namely SVM, LR, and DL methods, namely CNN and LSTM.

The feature extraction phase is a crucial step in SA that may influence the model's effectiveness. Among the most used techniques for word embedding, we found BERT, BERT (Devlin et al., 2019) is a model for representing language that is unsupervised and bidirectional. Unlike the context-independent glove and word2vec word embedding methods, BERT is a contextualized language model capable of producing various WE for the same term in different contexts. In various NLP tasks such as SA, BERT has obtained great achievements(Ji Min et al., 2023; Mann et al., 2023; Pota et al., 2020; Bensoltane and Zaki, 2022). used the Arabic version of BERT called AraBERT for Aspect-based sentiment analysis (ABSA). The outcomes demonstrate that BERT surpasses the other baseline techniques in the two tasks of ABSA, with more than 19% for the aspect category detection task and more than 6% for the aspect term extraction task. In the same context (Habbat et al., 2021), utilized AraBERT with a combination of BiLSTM and GRU to classify Arabic tweets, and the accuracy of the proposed model was 94%.

Numerous researchers have focused on the issue of classifying unbalanced data and have put forth numerous solutions. To deal with three unbalanced SA datasets (English and Turkish) utilized for binary categorization (Ogul and Guran, 2019), compared over-sampling and under-sampling techniques. As a result, it has been found that on datasets utilizing LR as a meta-classifier, the over-sampling techniques (SMOTE, Random OverSampling) increase the categorization model's effectiveness, while the under-sampling approaches (NearMiss and RUS) minimize the classification efficiency outcomes. However (Albahli, 2022), has developed a model that uses sentiment-based Twitter data for Gulf countries to identify genuine COVID-19-related news in Arabic text. The suggested SA model uses SMOTE and machine learning to handle unbalanced datasets. Unlike baseline models, SMOTE with Multinomial Naive Bayes provided the highest accuracy (91%). However (Roshan and Asadi, 2020), adopted the Bagging ensemble approach as one of the preferred techniques for ensemble learning because it is one of the few that, when combined with under-sampling techniques, can more effectively address imbalanced datasets.

Ensemble learning (EL) has recently become one of the most effective approaches. There are various ways to combine base classifiers, such as voting (Ardabili et al., 2019), bagging (Sagi and Rokach, 2018), averaging (Awni et al., 2019), boosting (Freund and Schapire, 1997), and stacking (Sarkar, 2020).

Most investigations demonstrate that EL techniques outperform single classifiers when the dataset is unbalanced. To create cost-sensitive models that account for misclassification (Chujai et al., 2015), proposed an approach for addressing the imbalanced classification of data problems using DT EL with bagging and boosting techniques. Compared to the Bag model, their results showed that boosting techniques, specifically AdaBoostM1, LogitBoost, RUSBoost, and TotalBoost, are the most efficient technique for identifying imbalanced data, particularly when there are overlapping categories and a large unbalance ratio (Tan et al.,

2022). combine the Robustly optimized Bidirectional Encoder Representations from the Transformers method (RoBERTa) and LSTM to classify English text using GloVe for data augmentation to solve the imbalance data issue. In the same context, the combination of RoBERTa, GRU, LSTM, and BiLSTM, together with the EL method, consists of averaging the ensemble and majority voting, is another example of a deep learning ensemble model that (Tan et al., 2022) presented. To help correct the imbalances in the datasets, pre-trained GloVe WE have also been implemented on the data. Their experimental results showed that ensemble models perform better when combined predictions, with both majority voting and averaging attaining higher levels of accuracy (89.81%). However (Muslim et al., 2023), proposed a new model to classify imbalanced datasets containing three optimization components; the first is SMOTE, an over-sampling method. The second aspect is the selection of features using LightGBM, and the third is stacking EL combining three base-learner classifiers (Random Forest, SVM, and KNN) into the meta-learner classifier XGBoost.

To summarize this part, as shown in Table 1, most of the examined studies opt for using an over-sampling method to balance the datasets and contextualized word embedding models to extract features. Some studies concluded that combining multiple deep learning models gives better results. Concerning the studied language, there are few works in languages like Arabic Dialect. Following examination, a variety of techniques for SA were used; the present analysis identifies the following key issues that are tackled in the implementation.

- Using different datasets in three languages: English, Standard Arabic, and Moroccan Arabic Dialect,
- Solving the imbalance class issue,
- Using contextualized word embedding (BERT for each language) to extract features,

-Using an ensemble stacking deep learning model to classify text.

3. The proposed approach

Stacking is a heterogeneous EL that utilizes a meta-classifier to combine several classification algorithms. A full training set is used for training the base classifiers, and the meta-classifier is trained using the outputs of the base-level model as features. Stacking is better than other methods because it can combine the best parts of different ML or DL algorithms to make accurate predictions greater than any single algorithm in the ensemble.

Based on stacked ensemble learning, we suggested a model for sentiment analysis on imbalanced datasets (in English, Arabic, and Moroccan Dialect) in this work. As seen in Fig. 1, the proposed method employs three DL models as base classifiers, namely LSTM, GRU, and BiLSTM, and that combines the outputs with a meta classifier. Using such stacked ensemble learning enables us to take advantage of each model's functional and structural advantages while improving overall achievement. In the following paragraphs, we will go over the base, meta classifiers, word embedding model (BERT), and data augmentation methods in greater detail.

The flow of our suggested model takes as input different imbalanced datasets in three languages: English, Arabic standard, and Moroccan Arabic Dialect; then, those datasets were balanced using SMOTE technique; after that, BERT was used to extract features, and finally, we classified the text into three classes using the stacking model composed of three base classifiers and SVM as a meta classifier.

3.1. Resampling datasets

There are two ways to create a balanced dataset from an unbalanced one: under-sampling and over-sampling.

3.1.1. Under-sampling

Undersampling balances a dataset by reducing the size of the plentiful class. This technique is utilized when sufficient data is available. By retaining all samples from the uncommon class and choosing an equivalent number of samples from the plentiful class, it is possible to obtain a new balanced dataset for modeling purposes. Our work used a random under-sampling (RUS) algorithm (Jian et al., 2016) as a user-sampling method.

3.1.2. Over-sampling

In contrast, oversampling is utilized when there is insufficient data. It attempts to balance the dataset by expanding the sample size of rare observations. Rather than deleting plentiful samples, new unusual samples are generated using repetition, bootstrapping, or SMOTE (Chawla et al., 2002), which we employed in our research.

3.2. BERT

BERT (Devlin et al., 2019) is an open-source machine learning technique for handling natural language (NLP). BERT employs the surrounding text to establish context in order to aid computers in interpreting confusing language. After being pre-trained on text from different sources, the BERT framework can be adjusted and fine-tuned with different tasks, including Sentiment analysis.

BERT is a Transformers-based deep learning model. Every output component is connected to each input element in Transformers, and their weightings are dynamically set according to their connection (attention mechanism). The following parts will define the BERT used for each language (English, Arabic Standard, and Moroccan Dialect).

3.2.1. BERT for English

BERT[1] (Bert-base-uncased, 2011) is a self-supervised pretrained on a large English corpus. This indicates that it was trained solely on the raw texts, without any human labeling; consequently, it may utilize a vast quantity of publicly accessible data, using an automated method to produce inputs and labels from the texts. More precisely, it was preconditioned with two goals.

-Masked language modeling (MLM): The technique hides 15% of the words in the input text at random, then processes the entire sentence to predict the hidden words.

- Next sentence prediction (NSP): The model enchains two masked sentences during pretraining. The model should then estimate whether or not the two sentences are consecutive.

The BERT was trained using BookCorpus, containing 11,038 unpublished books and English Wikipedia.

3.2.2. AraBERT for Arabic Standard

AraBERT[2] is an Arabic language model pretrained using Google's BERT architecture. Arabic ELECTRA and GPT2 also utilize the pretraining data utilized by the new AraBERT model. The dataset before applying Farasa Segmentation was 77 GB (8,655,948,860 words or 82,232,988,358 characters) in length.

The AraBERT models were evaluated on various downstream tasks and compared with mBERT and other cutting-edge models. The tasks consisted of sentiment analysis on six distinct datasets.

3.2.3. DarijaBERT for Moroccan Dialect

DarijaBERT[3] (Kamel/DarijaBERT, 2022) is the first Open Source BERT model for Moroccan Arabic's "Darija" dialect. AIOX Labs, an AI startup headquartered in Rabat, Morocco, designed and made available this language model. DarijaBERT employs BERT-base's architecture,

[1] https://huggingface.co/bert-base-uncased.
[2] https://huggingface.co/aubmindlab/bert-base-arabert.
[3] https://huggingface.co/SI2M-Lab/DarijaBERT.

Table 1
Summary of related work.

Ref	Dataset	Resampling techniques for Imabalanced data	Word emebdding techniques	Classification techniques	Results
Hicham et al. (2022)	3 Arabic datasets	–	TF-IDF	-Adaboost classifier, -SVM, Maximum Entropy, -Decision tree , KNN -Ensemble machine learning	Accuracy >87,9% for ensemble machine learning model.
M.Abdelgwad et al. (2021)	Arabic dataset	–	Word2vec FastText	CNN LSTM IAN IAN-CNN IAN-LSTM IAN-BiLSTM IAN-GRU IAN-BiGRU	Accuracy of 83,90% for IAN + BiGRU using FastText
Abu Kwaik et al. (2019)	Dialectal Arabic	–	AraVec	-LSTM + LSTM -BiLSTM + LSTM -BiLSTM + BiLSTM -BiLSTM + CNN	BiLSTM + CNN is better
Habbat et al. (2022)	French dataset	–	-XLNet -CamemBERT -MultiFiT	CNN LSTM BiLSTM GRU LSTM + CNN BiLSTM + CNN GRU + CNN	Accuracy >89.6% for GRU + CNN using XLNet
Garouani and Kharroubi (2022b)	-Standard Arabic dataset -Arabic Dialect dataset	–		-LR -SVM -CNN -LSTM	Accuracy >90% for LSTM
Bensoltane and Zaki (2022)	Arabic dataset	–	AraBERT	-BERT + BiLSTM + CRF -BERT + CRF -BERT + Linear -BERT + BiGRU + CRF	Precision of 87,7% for BERT + BiGRU + CRF
Habbat et al. (2021)	Arabic dataset	–	-AraVEC -FastText -AraBERT	-CNN -LSTM -BiLSTM -GRU -BiGRU -BiLSTM + GRU	Accuracy of 94% for BiLSTM + GRU using AraBERT
Ogul and Guran (2019)	English dataset Turkish Datset	RUS NM SMOTE ROS	-Word-based N-gram		ROS, SMOTE increases the performance
Albahli (2022)	Arabic dataset	SMOTE	BoW	Naïve Bayes	SMOTE + Naïve Bayes
Roshan and Asadi (2020)		Bagging method		Bagging method	Bagging is better to solve imbalance class issues
Chujai et al. (2015)		Bagging		AdaBoostM1 Bag TotalBoost LogitBoost RUSBoost	Boosting is better (RUSBoost, TotalBoost, LogitBoost, and AdaBoostM1
Tan et al. (2022b)	English datasets	GloVe	RoBERTa	-LSTM -BiLSTM -GRU -Averaging ensemble method -Voting ensemble method	Accuracy of 89.81% for ensemble methods
Muslim et al. (2023)	Online P2P Lending dataset	SMOTE	LightGBM	-KNN -SVM -Random forest -Ensemble model (KNN, SVM, random forest)	Ensemble model is better
Our proposed model	-English dataset, -Arabic datasets, -Dialect datasets	-SMOTE -RUS	BERT	-CNN -LSTM -BiLSTM -GRU -Ensemble stacking model	The proposed model is better.

excluding the Next Sentence Prediction (NSP) target. There are around 3 Million Darija sequences containing 691 MB of text or a total of 100M tokens that were used to train this model. No publication describing the training was available.

The model was trained on data from three various sources.

1. Stories in Darija extracted from a specialized website
2. YouTube reviews from forty distinct Moroccan channels
3. Tweets that have been collected based on a set of Darija keywords.

The maximum input length for this language model is 128, and the vocabulary size is 80.000 tokens.

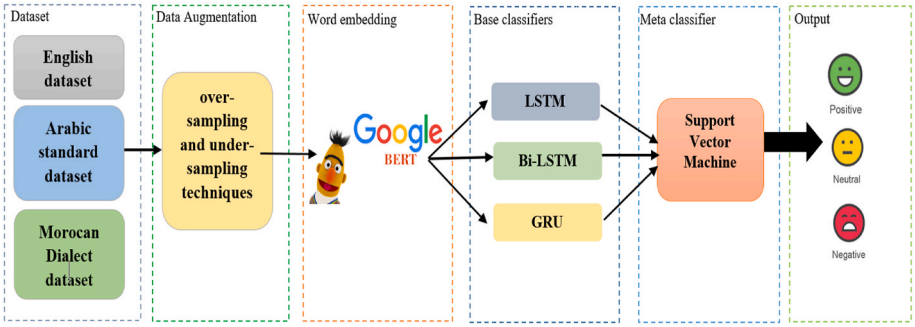

Fig. 1. The general structure.

3.3. The base classifiers and meta-classifier

This part will define the base classifiers, LSTM, Bi-LSTM, and GRU, and the meta-classifier SVM.

3.3.1. LSTM

LSTM is a type of RNN (Sherstinsky, 2020); it addresses the vanishing gradients issue. It involves a memory cell and 3 gates: the input, forget, and output gates which regulate the movement of information across the network. The memory cell keeps the gradients unchanged if the forget gate is activated and the input and output gates are turned off. As a result, LSTM minimizes the vanishing gradient issue and can learn long-term dependencies.

3.3.2. Bi-LSTM

The bidirectional LSTM model is an RNN that comprises 2 LSTMs, an LSTM taking input forward and the other backward. As a result, it obtains more information from the text by understanding what text follows and precedes words in sentences by analyzing the entire sequence in both directions (G. Q. Xu et al., 2019).

3.3.3. GRU

The GRU model is like the LSTM, except it has fewer settings making it faster than LSTM. In contrast to LSTM, the GRU has just 2 gates: the reset and update gates. The update gate manages the flow of information into memory, while the reset gate regulates the flow of information out of memory. If the update gate is set to 1, the last memory is fully preserved; otherwise, the last memory is completely erased (update gate = 0). GRU is faster than LSTM, but its performance depends on the dataset characteristics.

3.3.4. SVM

SVMs are supervised learning techniques for classification, regression, and detecting outliers. The benefits of SVM are as follows.

- ✓ Effective in spaces with high dimensions.
- ✓ Still efficient in situations where the number of dimensions exceeds the number of samples.
- ✓ It uses a subset of training points (called support vectors) in the function of decision, making it memory effective.
- ✓ Various Kernel functions may be specified for the decision function, making it versatile. Standard kernels are available, but it is also possible to specify custom kernels.

Our proposal is based on a stacking ensemble method. After training our base classifiers, we used the stacking method to combine their results. In this paper, we combine the results of the above classifiers with the help of SVM. The composed layers of each base classifier are presented in the Table below.

Concerning the implementation, we used Python Language (Pytorch library), and as the configuration of our model, we used 25 epochs and the batch size equal to 64 for training (Tested values are presented in the experimental findings part).

4. Results and experiments analysis

This part outlines the datasets, the evaluation metrics, and the analysis of the results, including applying the best model to the use case concerning ONCF comments.

4.1. Datasets

We employ our suggested model to real English, Arabic, and Moroccan Dialect datasets to test it. The used datasets are summarized in Table 2.

4.1.1. English datasets

The English datasets used in this work are the Sentiment140 dataset and the Twitter US Airline Sentiment dataset.

- Sentiment140: Sentiment140 (Go et al., 2022) is a large dataset obtained from Twitter by Stanford University. It includes 1,600,000 tweets that were collected using the Twitter API.
- Twitter US Airline Sentiment: CrowdFlower gathered the 2017 Twitter US Airline Sentiment dataset (Twitter US Airline Sentiment, 2011) from Twitter. Customer feedback is included for six American airlines in the dataset: American, US Airways, United, South-West, Virgin America Airlines, and Delta. Positive, negative, and neutral sentiment classes are represented in this dataset by 2363, 9178, and 3099 samples, respectively.

4.1.2. Arabic datasets

The Arabic datasets used in this study are the Arabic-Reviews of Hotels Dataset and the Arabic Twitter dataset.

- Arabic-Reviews of Hotels Dataset (HARD): This dataset (Elnagar et al., 2018) includes 93,700 hotel reviews in Arabic. The hotel comments were gathered from the Booking.com website in June and July 2016. Dialectal Arabic, and Modern Standard Arabic are used in the comments. Our study used an unbalanced dataset, which contains 373,750 reviews.
- Arabic Twitter dataset: The dataset (Albahli, 2022) included Arabic tweets shared during the COVID-19 pandemic. The data were only available from the Gulf countries of Qatar, Oman, Bahrain, Saudi

Table 2
Architecture of each base classifier.

Classifier	Architecture
LSTM	Embedding layer → LSTM layer → Dense layer
Bi-LSTM	Embedding layer → Bi-LSTM layer → Dense layer
GRU	Embedding layer → GRU layer → Dense layer

Arabia and the United Arab Emirates (UAE). The dataset was scraped from Twitter between March 2020 and April 2020. The keywords coronavirus, corona, covid19, sarscov2, and COVID were used to collect 60,000 tweets.

4.1.3. Moroccan Dialect datasets

Our model was tested on two datasets: the Moroccan Arabic Corpus (MAC) (Garouani and Kharroubi, 2022a) and the Dialect dataset (MSDA, 2022). The characteristics of the datasets utilized to evaluate the suggested approach are summarized in Table 3, and label examples for each used dataset are presented in Table 4.

-MAC dataset: The MAC dataset is the largest and first publicly available Moroccan Arabic Corpus for SA. It is a pioneer in size, quality, and availability to the academic research community, as evidenced by the native annotators' consistency (IAA = 0.9). This large and free MAC comprised 18000 manually classified tweets, yielding a lexical dictionary of 30000 negative, positive, and neutral words.

-Dialect Dataset: This dataset is one of the MSDA open datasets, and it contains a collection of about 50K tweets that have been classified by country-level dialect. Because all five dialects in this dataset have Arabic roots, many parallels exist, especially among geographically close ones, such as Moroccan and Algerian dialects.

4.2. Evaluation metrics

To determine the method's efficacy that we have presented, we have applied it to the two datasets above-presented and compared the results with those obtained by base classifiers (LSTM, BiLSTM, GRU, and CNN) utilizing Accuracy, Mattheus Correlation Coefficient (MCC), ROC-AUC curve, and Cohen's Kappa, as performance measures.

4.2.1. Accuracy

Accuracy is one of the most widely used metrics in multi-class classification, and it simply reports the proportion of properly categorized phrases, irrespective of their class.

$$Acc = \frac{Number\ of\ correct\ classification}{Total}$$

4.2.2. Matthew's correlation coefficient

Matthew's correlation coefficient, commonly called MCC, was first developed by Brian Matthews in 1975 (Matthews, 1975). A statistical measure is used to evaluate models. Its purpose is to evaluate or quantify the disparity between expected and actual values, and it is analogous to chi-square statistics when applied to a contingency table with two rows and two columns.

MCC only became widely used in the 2000s to evaluate the performance of ML and DL techniques with some modifications to the multi-class case. Matthew's correlation coefficient is the best single-value classification measurement for summarizing an error matrix or confusion. MCC is computed using the following formula:

$$MCC = \frac{TN * TP - FN * FP}{\sqrt{(FP + TP)(TP + FN)(FP + TN)(TN + FN)}}$$

Table 3
Datasets characteristics.

Language	Dataset	Positive	Negative	Neutral	Total
English	**Sentiment140**	3000	7500	–	10500
	US Airline	2363	9178	3099	14640
Arabic	**HARD**	18300	5490	6710	30500
	Arabic Twitter	6690	4014	11596	22300
Moroccan Dialect	**MAC dataset**	9000	2000	4000	15000
	Dialect dataset	4500	2100	7600	13300

A few changes occur when it applies to the multi-class classification: the True and Estimated class distributions are no longer binary, and a greater number of classes are taken into consideration. In this situation, the numerator and denominator have a different shape than in the binary case, which can help to identify more consistent results within MCC's range [−1; +1].

$$MCC = \frac{c * s - \sum_{k}^{K} p_k * t_k}{\sqrt{\left(s^2 - \sum_{k}^{K} p_k^2\right)\left(s^2 - \sum_{k}^{K} t_k^2\right)}}$$

Where : The number of properly predicted elements: $c = \sum_{k}^{K} C_{kk}$.

The number of elements : $s = \sum_{i}^{K}\sum_{j}^{K} C_{ij}$.

The number of times class k has been predicted: $p_k = \sum_{i}^{K} C_{ki}$.

The number of times class k did really happen: $t_k = \sum_{i}^{K} C_{ik}$.

4.2.3. Cohen's kappa

The kappa score (Cohen, 1960) is a supercharged version of accuracy including chance and class imbalance measurements. It is a statistic utilized to evaluate the inter-and intra-rater reliability of qualitative items. Generally, it is considered a more reliable metric than a simple percent agreement computation because it considers the possibility that the agreement could have occurred by chance. It has the following formula:

$$Kappa = \frac{p_0 - p_e}{1 - p_e}$$

Where p_0 is the overall accuracy of the model and p_e is the degree to which model predictions and actual class values accord as though by accident.

The calculation of Cohen's Kappa Score adjusts structure in the multi-class situation, becoming more similar to MCC, and its formula is as follows:

$$Kappa = \frac{c * s - \sum_{k}^{K} p_k * t_k}{s^2 - \sum_{k}^{K} p_k * t_k}$$

4.2.4. ROC-AUC

The AUC - ROC curve, also known as the Area Under the Receiver Operating Characteristics curve, is a measure of performance that can be applied to classification problems at various threshold levels. The Area under the curve (AUC) is the degree of separability, while the receiver operating characteristic (ROC) curve is a probability curve. It indicates how well the model can differentiate between different types of data.

4.3. Experimental findings

This part aims to compare the effectiveness of the suggested method outlined in section 3 to that of various DL models, namely LSTM, CNN, Bi-LSTM, and GRU, using BERT as the word embedding model and SMOTE and RUS as resampling dataset techniques. We used six datasets (2 for each language: English, Arabic Standard, and Moroccan Dialect) to test our model.

Referring to Table 5, on the Semtiment140 dataset with up to 10,000 tweets between positive and negative, the proposed model using SMOTE achieved an accuracy of 94% with k-fold validation = 10 and 93.9% with k-fold validation = 5. On the US Airline dataset with approximately 14,700 English reviews between positive, negative, and neutral, the proposed model has a higher accuracy of 93.9% with k-fold validation = 10 and 93.7% with k-fld validation = 5.

As observed in Table 6 concerning Arabic datasets, on the Arabic Twitter dataset with 22,300 tweets, the suggested model using SMOTE

Table 4
Label examples for used datasets.

Language	Dataset	Positive	Negative	Neutral
English	**Sentiment140**	Dont worry too much your friends at twitter will be supporting you	my whole body feels itchy and like its on fire	It's okey
	US Airline	virginmedia I'm flying your #fabulous #Seductive skies again! U take all the #stress away from travel	VirginAmerica you guys messed up my seating. I reserved seating with my friends and you guys gave my seat away	VirginAmerica Really missed a prime opportunity for Men Without Hats parody
Arabic	**HARD**	ممتاز". النظافة والطاقم متعاون".	غرفة قياسية مزدوجة أقمت ليلة واحدة "استغرب تقييم الفندق كخمس نجوم". لا شي. يستحق 2 نجمه	لاشي"
	Arabic Twitter	الحياة أجمل من حزن و أصدق من خيبة و أعظم من غدر و أكبر من كذبة. فقط. أنظر لها بـ القلب الذي يرى الله. و ثق به حتى إن ك	نصمت !! تسير حياتنا على مَ يرام فالنّاس لم تعد كما كانت	لا نخلو من ضغوطات الحياة. فنحن نعيش على أرض أعدت للبلاء ولم يسلم منها حتى الأنبياء
Moroccan Dialect	**MAC dataset**	هادو كيبينو انهم ماشي غير كيشجعو فرقة دالكورة هادو ناس يترفع ليهم الشابو	هجومك يا أخي على حسن طارق تبرهيش و قلة عقل ودليل انعدام غرام ديال المنطق	عاد إليكم من جديد وأخيرا درت جيم لهاد الصفحة
	Dialect dataset	الله يعوضك خير ان شاء الله	هاد المرحلة صعيبيبة ديال يكون عندك ومن بعد تخسر مالك , خاصة على لوليدات الصغار لي كيكونو تزاد فالنور ومعرفوش معنى الظلا	ها علاش مكنقراوش للقصص للاطفال قبل النوم

Table 5
Comparison of various performance measures for different models on English datasets with k-fold validation.

Model	Resampling method	Sentiment140 dataset			US Airline Dataset		
		Acc	MCC	Kappa	Acc	MCC	Kappa
K-fold validation (K=5)							
CNN	**RUS**	0.898	0.780	0.719	0.885	0.785	0.700
	SMOTE	0.900	0.800	0.788	0.901	0.797	0.773
LSTM	**RUS**	0.900	0.794	0.724	0.890	0.790	0.712
	SMOTE	0.917	0.810	0.791	0.910	0.801	0.781
Bi-LSTM	**RUS**	0.912	0.797	0.787	0.900	0.898	0.872
	SMOTE	0.919	0.820	0.798	0.912	0.815	0.790
GRU	**RUS**	0.918	0.877	0.883	0.908	0.872	0.851
	SMOTE	0.923	0.845	0.809	0.919	0.818	0.804
The proposed model	**RUS**	0.930	0.875	0.858	0.922	0.896	0.885
	SMOTE	0.939	0.885	0.823	0.937	0.877	0.813
K-fold validation (K = 10)							
CNN	**RUS**	0.899	0.781	0.720	0.887	0.787	0.702
	SMOTE	0.901	0.802	0.789	0.902	0.798	0.774
LSTM	**RUS**	0.902	0.795	0.725	0.892	0.791	0.716
	SMOTE	0.919	0.811	0.793	0.911	0.803	0.783
Bi-LSTM	**RUS**	0.911	0.799	0.789	0.902	0.898	0.874
	SMOTE	0.920	0.821	0.799	0.913	0.817	0.792
GRU	**RUS**	0.919	0.878	0.885	0.910	0.875	0.854
	SMOTE	0.926	0.846	0.810	0.921	0.820	0.806
The proposed model	**RUS**	0.932	0.877	0.860	0.927	0.897	0.887
	SMOTE	0.940	0.886	0.824	0.939	0.879	0.815

method yields a best accuracy of 94% with k-fold validation = 10 and 93.9% with k-fold validation = 5. On the HARD dataset with 30,500 Arabic hotel reviews, the presented model outperforms the baseline models with an accuracy of 94,2% with k-fold validation = 10 and 94% with k-fold validation = 5.

As shown in Table 7 concerning the Moroccan dialect datasets, the proposed model achieved 94.3% and 94.2% accuracy on the MAC dataset with k-fold validation of 10 and 5, respectively. On the Dialect dataset, the suggested model yields 94.1% and 94% accuracy with k-fold validation of 10 and 5, respectively.

In summary, Tables 5–7 compare the proposed approach and different DL algorithms on each dataset using RUS and SMOTE as resampling methods regarding accuracy, MCC, and Cohen's Kappa with 5-fold validation and 10-fold validation. Clearly, the suggested approach improved the performance of SA, as it achieved an accuracy of more than 94% on different datasets using BERT as word embeddings and 10-fold validation.

Concerning the re-sampling methods, It has been discovered that the over_sampling method SMOTE increases the performance values of the classifier. In contrast, the under-sampling method RUS decreases the performance values of the datasets.

In brief, using the over_sampling technique, the BERT word embedding model, and the stacking model classification allows better accuracy. This suggests that the proposed architecture influences sentiment classification on three dependent-language datasets. Instead, results obtained for the single classifiers are better using GRU than Bi-LSTM or LSTM, and CNN comes last. In particular, better accuracy and a comparable MCC and Cohen's kappa for stacking model can be achieved. In addition to these results, an ablation study was performed regarding ROC_AUC, the memory/time of execution. The results of the ablation study are presented in the following Figures and Tables.

The ROC score for each baseline Deep Learning model on each dataset used in the experiments is calculated in Figs. 2–4. As demonstrated, the ROC scores of single classifiers, especially CNN, are the lowest in the various datasets when compared to other deep learning models and stacked deep learning model in particular.

The comparison of each model's memory usage on each dataset is shown in Figs. 5–7. It appears that the CNN model requires greater computing power, and dealing with enormous amounts of data will exponentially increase the computing burden. Thus, the proposed model needs more memory than the single RNN models. Despite the model's good performance in terms of the metrics used, the runtime is a little long (average of 540 s) as shown in Table 8, so the proposed model remains to be improved to have the best performance in terms of runtime and memory.

In addition, we did the performance metrics for all models when

Table 6
Comparison of various performance measures for different models on Arabic datasets with k-fold validation.

Model	Resampling method	HARD dataset			Arabic Twitter Dataset		
		Acc	MCC	Kappa	Acc	MCC	Kappa
K-fold validation (K=5)							
CNN	**RUS**	0.898	0.786	0.710	0.895	0.775	0.703
	SMOTE	0.900	0.800	0.773	0.918	0.802	0.772
LSTM	**RUS**	0.898	0.790	0.719	0.900	0.787	0.713
	SMOTE	0.915	0.812	0.790	0.922	0.811	0.788
Bi-LSTM	**RUS**	0.907	0.785	0.783	0.898	0.897	0.870
	SMOTE	0.917	0.822	0.787	0.928	0.817	0.792
GRU	**RUS**	0.915	0.867	0.885	0.895	0.869	0.848
	SMOTE	0.923	0.843	0.812	0.935	0.820	0.801
The proposed model	**RUS**	0.930	0.869	0.864	0.932	0.895	0.882
	SMOTE	0.940	0.882	0.825	0.939	0.879	0.816
K-fold validation (K = 10)							
CNN	**RUS**	0.899	0.787	0.712	0.897	0.778	0.705
	SMOTE	0.901	0.802	0.774	0.920	0.803	0.775
LSTM	**RUS**	0.901	0.791	0.720	0.902	0.789	0.714
	SMOTE	0.917	0.815	0.791	0.927	0.813	0.789
Bi-LSTM	**RUS**	0.909	0.787	0.784	0.900	0.899	0.871
	SMOTE	0.919	0.823	0.789	0.930	0.819	0.794
GRU	**RUS**	0.918	0.868	0.887	0.897	0.871	0.850
	SMOTE	0.924	0.844	0.814	0.938	0.821	0.802
The proposed model	**RUS**	0.931	0.870	0.865	0.934	0.897	0.884
	SMOTE	0.942	0.885	0.827	0.940	0.880	0.817

Table 7
Comparison of various performance measures for different models on Moroccan Dialect datasets with k-fold validation.

Model	Resampling method	MAC dataset			Dialect Dataset		
		Acc	MCC	Kappa	Acc	MCC	Kappa
K-fold validation (K=5)							
CNN	**RUS**	0.895	0.776	0.711	0.873	0.770	0.702
	SMOTE	0.901	0.805	0.770	0.906	0.809	0.750
LSTM	**RUS**	0.900	0.788	0.729	0.889	0.788	0.721
	SMOTE	0.915	0.824	0.781	0.912	0.811	0.785
Bi-LSTM	**RUS**	0.917	0.787	0.780	0.913	0.896	0.879
	SMOTE	0.924	0.823	0.789	0.924	0.812	0.787
GRU	**RUS**	0.919	0.870	0.888	0.916	0.867	0.885
	SMOTE	0.937	0.845	0.890	0.938	0.840	0.841
The proposed model	**RUS**	0.935	0.870	0.883	0.937	0.867	0.880
	SMOTE	0.942	0.884	0.889	0.940	0.880	0.878
K-fold validation (K = 10)							
CNN	**RUS**	0.897	0.778	0.712	0.875	0.772	0.704
	SMOTE	0.902	0.806	0.772	0.908	0.810	0.752
LSTM	**RUS**	0.901	0.789	0.730	0.900	0.787	0.724
	SMOTE	0.917	0.825	0.782	0.913	0.812	0.787
Bi-LSTM	**RUS**	0.919	0.789	0.781	0.914	0.897	0.880
	SMOTE	0.927	0.824	0.790	0.926	0.815	0.789
GRU	**RUS**	0.920	0.871	0.890	0.918	0.869	0.887
	SMOTE	0.939	0.847	0.891	0.940	0.841	0.842
The proposed model	**RUS**	0.937	0.871	0.884	0.938	0.869	0.881
	SMOTE	0.943	0.885	0.890	0.941	0.881	0.880

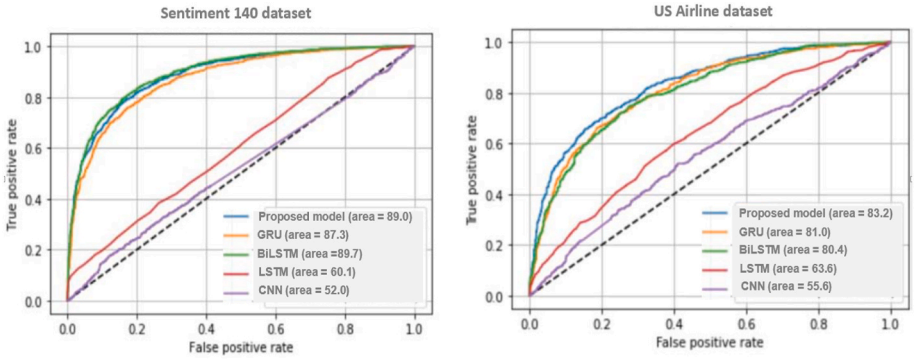

Fig. 2. The ROC-AUC curve on the English datasets using SMOTE technique.

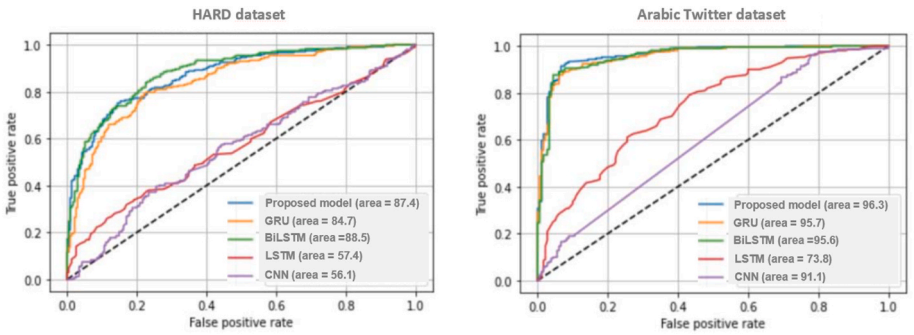

Fig. 3. The ROC-AUC curve on the Arabic datasets using SMOTE technique.

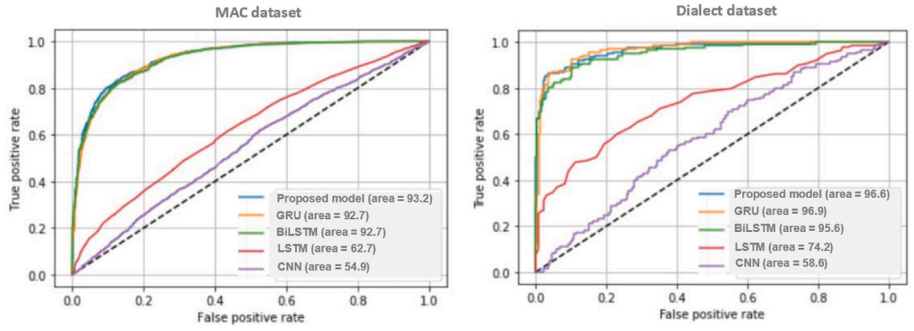

Fig. 4. The ROC-AUC curve on the Moroccan Dialect datasets using SMOTE technique.

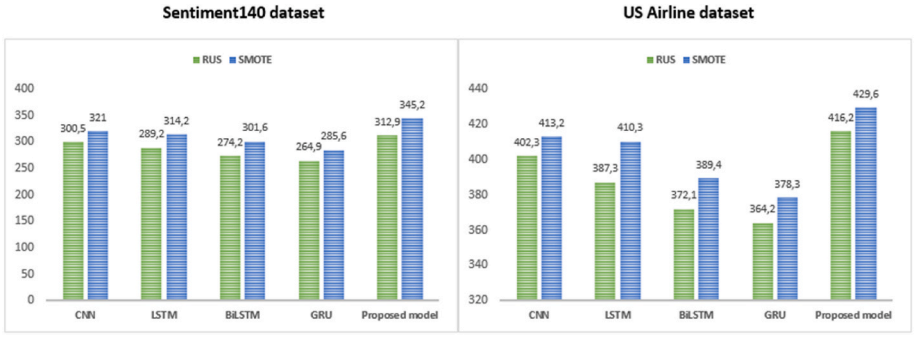

Fig. 5. The memory usage (in MiB) of different models on the English datasets.

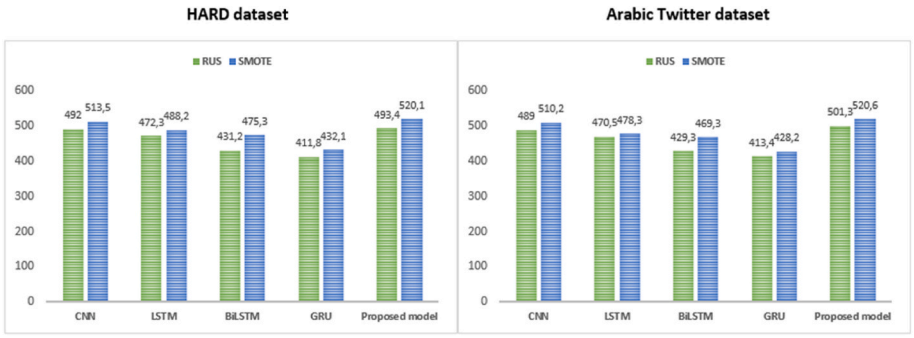

Fig. 6. The memory usage (in MiB) of different models on the Arabic datasets.

scaling the dataset for training and testing like 90:10, 80:20, 70:30, and 60:40. As shown in Tables 9–11, the scaling 70:30 has the best performance in different datasets. Also, we tested different hyper-parameters to find the optimal ones to run our proposed model, as shown in Table 12.

After evaluating our proposed approach, we applied the model to a collected dataset using web scraping techniques; this dataset consists of 19,786 Darija comments published on the ONCF Facebook page. As shown in Fig. 8, the Moroccan users gave more positive comments concerning ONCF services, with 63.9%.

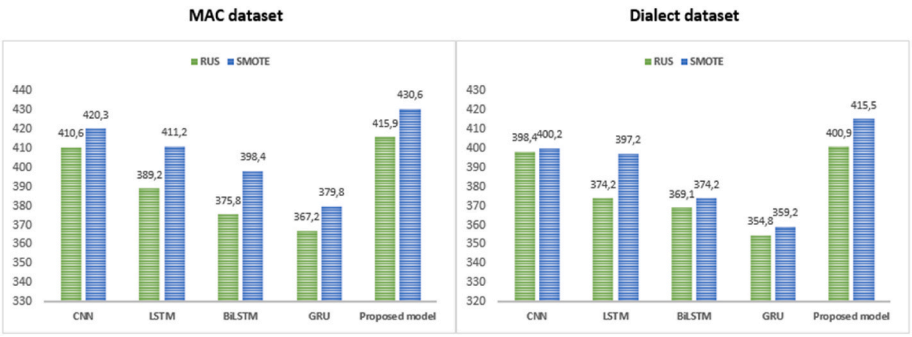

Fig. 7. The memory usage (in MiB) of different models on the Dialect datasets.

Table 8
Average of Training/prediction time for different models.

Model	Resampling techniques	Average of Training/prediction time
CNN	RUS	463 s
	SMOTE	475 s
LSTM	RUS	448 s
	SMOTE	457 s
BiLSTM	RUS	426 s
	SMOTE	438 s
GRU	RUS	416 s
	SMOTE	422 s
The proposed model	RUS	530 s
	SMOTE	540 s

5. Conclusion and future work

The researcher has recently been interested in Arabic dialect sentiment analysis. This paper suggested a model for SA of the Moroccan dialect as well as English and Arabic standard. The suggested framework is an ensemble learning-based deep neural network architecture. LSTM, BiLSTM, and GRU were created and trained before in the proposed approach. The outputs of base classifiers are then combined using an SVM as a meta-classifier. The used datasets are imbalanced, so we compared the under-sampling (RUS method) and over-sampling (SMOTE method) for resampling datasets. As a result, we found that SMOTE outperforms the RUS method in all datasets, and the stacking model achieves the best performance using BERT as a word embedding model with 10-fold validation.

We implemented our technique using real datasets. The evaluation results demonstrate that the proposed approach is more accurate than the baseline models in terms of Cohen's kappa, accuracy, and MCC despite consuming more prediction memory/time, and the latter, given that the model uses the SMOTE method, which allows adding more text to the minor class and the use of the stacking model which combines different classifiers.

Since the outcome of the ensemble technique is promising, we intend to implement ensemble learning with different meta-classifiers using several word embedding models in a future study. Also, we aim to use a Generative pre-trained transformer (GPT-3) for data augmentation. GPT -3 is a deep learning-based autoregressive language model released in 2020. When given a prompt, it will generate text that continues the

Table 9
Accuracy for different models on English datasets when scaling the datasets for training and testing.

Model	Resampling technqiues	60:40		70:30		80:20		90:10	
Datasets		Sentiment 140	US Airline	Sentiment 140	US Airline	Sentiment 140	US Airline	Sentiment 140	US Airline
CNN	***RUS***	0.897	0.886	0.899	0.887	0.898	0.886	0.895	0.885
	SMOTE	0.899	0.900	0.901	0.902	0.900	0.900	0.889	0.900
LSTM	***RUS***	0.900	0.891	0.902	0.892	0.898	0.889	0.887	0.888
	SMOTE	0.910	0.914	0.919	0.911	0.912	0.908	0.900	0.907
BiLSTM	***RUS***	0.909	0.900	0.911	0.902	0.903	0.888	0.901	0.889
	SMOTE	0.915	0.907	0.920	0.913	0.916	0.902	0.900	0.887
GRU	***RUS***	0.917	0.902	0.919	0.910	0.911	0.900	0.907	0.888
	SMOTE	0.920	0.914	0.926	0.921	0.919	0.905	0.915	0.902
Proposed model	***RUS***	0.930	0.921	0.932	0.927	0.929	0.920	0.928	0.918
	SMOTE	0.938	0.935	0.940	0.939	0.937	0.930	0.933	0.927

Table 10
Accuracy for different models on Arabic datasets when scaling the datasets for training and testing.

Model	Resampling technqiues	60:40		70:30		80:20		90:10	
Datasets		HARD	Twitter	HARD	Twitter	HARD	Twitter	HARD	Twitter
CNN	***RUS***	0.898	0.896	0.899	0.897	0.897	0.895	0.897	0.895
	SMOTE	0.900	0.918	0.901	0.920	0.900	0.917	0.901	0.900
LSTM	***RUS***	0.989	0.900	0.901	0.902	0.888	0.889	0.889	0.887
	SMOTE	0.915	0.925	0.917	0.927	0.911	0.921	0.910	0.908
BiLSTM	***RUS***	0.907	0.889	0.909	0.900	0.908	0.887	0.908	0.907
	SMOTE	0.915	0.927	0.919	0.930	0.914	0.925	0.913	0.924
GRU	***RUS***	0.912	0.890	0.918	0.897	0.913	0.989	0.912	0.987
	SMOTE	0.920	0.937	0.924	0.938	0.919	0.935	0.918	0.934
Proposed model	***RUS***	0.927	0.931	0.931	0.934	0.926	0.930	0.924	0.937
	SMOTE	0.940	0.939	0.942	0.940	0.939	0.938	0.938	0.937

Table 11
Accuracy for different models on Dialect datasets when scaling the datasets for training and testing.

Model	Resampling technqiues	60:40		70:30		80:20		90:10	
Datasets		MAC	Dialect	MAC	Dialect	MAC	Dialect	MAC	Dialect
CNN	***RUS***	0.896	0.876	0.897	0.875	0.895	0.877	0.895	0.876
	SMOTE	0.900	0.900	0.902	0.908	0.889	0.889	0.888	0.887
LSTM	***RUS***	0.901	0.887	0.901	0.900	0.900	0.886	0.901	0.887
	SMOTE	0.915	0.912	0.917	0.913	0.914	0.911	0.913	0.912
BiLSTM	***RUS***	0.917	0.913	0.919	0.914	0.915	0.912	0.914	0.911
	SMOTE	0.926	0.925	0.927	0.926	0.924	0.924	0.924	0.921
GRU	***RUS***	0.919	0.916	0.920	0.918	0.911	0.915	0.910	0.912
	SMOTE	0.937	0.938	0.939	0.940	0.930	0.939	0.929	0.937
Proposed model	***RUS***	0.936	0.937	0.937	0.938	0.937	0.936	0.930	0.935
	SMOTE	0.942	0.940	0.943	0.941	0.941	0.938	0.940	0.937

Table 12
Experimental parameters.

The hyper-parameter	Score ranges	Best value
Epoch	30.15. 20, 25	25
Batch size	80. 32. 64. 8.16	64
Optimizer	Adadelta, Adagrad, Adamax, Adam, Adadelta, Nadam	Adam
Activation function	softmax, softplus, relu, tanh, linear	softmax
Dropout	0,6. 0,5. 0,4. 0,2	0,4

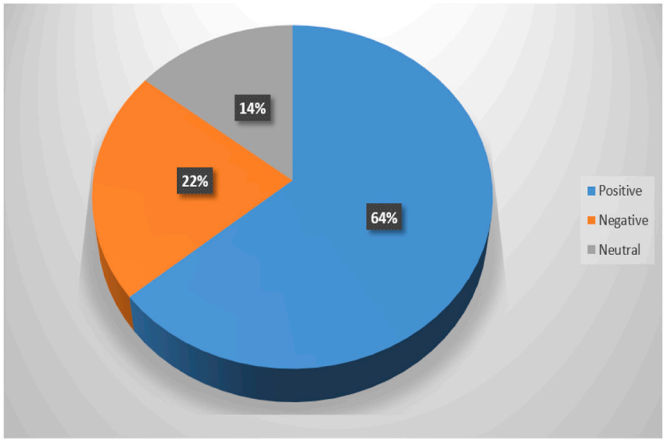

Fig. 8. Sentiment analysis on ONCF comments using the proposed model.

prompt; therefore, we can use it to generate realistic samples and improve the classification task. To speed up the process, we will implement it on the Apache spark framework, which allows data analysis in parallel.

CRediT authorship contribution statement

Nassera Habbat: Conceptualization, Methodology, Resources, Software, Data curation, Writing – original draft, Making the Revisions. **Hicham Nouri:** Resources, Visualization, Investigation, Data curation, Writing – original draft, Preparation of datasets. **Houda Anoun:** Conceptualization, Methodology, Supervision, Validation, Check the revisions. **Larbi Hassouni:** Supervision, Validation, Check the revisions.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

I have mentioned the GitHub link in the \"Detailed response to reviewers\" file and uploaded the used datasets in the \"Attach file\" step as supplementary files.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.engappai.2023.106999.

References

Abdelgwad, M., Soliman, M.A., I Taloba A, T.H., Farghaly, M.F., 2021. Arabic Aspect Based Sentiment Analysis Using Bidirectional GRU Based Models. Journal of King Saud University - Computer and Information Sciences. https://doi.org/10.1016/j.jksuci.2021.08.030. S1319157821002482.

Abu Kwaik, K., Saad, M., Chatzikyriakidis, S., Dobnik, S., 2019. LSTM-CNN deep learning model for sentiment analysis of dialectal Arabic. In: Smaïli, K. (Ed.), Arabic Language Processing : from Theory to Practice, 1108. Springer International Publishing, pp. 108–121. https://doi.org/10.1007/978-3-030-32959-4_8.

Alali, M., Mohd Sharef, N., Azmi Murad, M.A., Hamdan, H., Husin, N.A., 2022. Multitasking learning model based on hierarchical attention network for Arabic sentiment analysis classification. Electronics 11 (8), 1193. https://doi.org/10.3390/electronics11081193.

Albahli, S., 2022. Twitter sentiment analysis : an Arabic text mining approach based on COVID-19. Front. Public Health 10, 966779. https://doi.org/10.3389/fpubh.2022.966779.

Ardabili, S., Mosavi, A., Várkonyi-Kóczy, A.R., 2019. *Advances in machine learning modeling reviewing Hybrid and ensemble methods* [preprint]. MATHEMATICS & COMPUTER SCIENCE. https://doi.org/10.20944/preprints201908.0203.v1.

Awni, M., Khalil, M.I., Abbas, H.M., 2019. Deep-Learning Ensemble for Offline Arabic Handwritten Words Recognition. *2019 14th International Conference on Computer Engineering and Systems*, 40–45. ICCES). https://doi.org/10.1109/ICCES48960.2019.9068184.

Bensoltane, R., Zaki, T., 2022. Towards Arabic aspect-based sentiment analysis : a transfer learning-based approach. Social Network Analysis and Mining 12 (1), 7. https://doi.org/10.1007/s13278-021-00794-4.

Bert-base-uncased · Hugging Face. (2011). Consulté 24 janvier 2023, à l'adresse https://huggingface.co/bert-base-uncased.

Boujou, E., Chataoui, H., Mekki, A.E., Benjelloun, S., Chairi, I., Berrada, I., 2021. An Open Access NLP Dataset for Arabic Dialects : Data Collection, Labeling, and Model Construction arXiv:2102.11000). arXiv. http://arxiv.org/abs/2102.11000.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE : synthetic minority over-sampling technique. J. Artif. Intell. Res. 16, 321–357. https://doi.org/10.1613/jair.953.

Chujai, P., Chomboon, K., Teerarassamee, P., Kerdprasop, N., Kerdprasop, K., 2015. Ensemble learning for imbalanced data classification problem. The Proceedings of the 2nd International Conference on Industrial Application Engineering 2015, 449–456. https://doi.org/10.12792/iciae2015.079.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20 (1), 37–46. https://doi.org/10.1177/001316446002000104.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2019. BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv:1810.04805 [Cs]*. http://arxiv.org/abs/1810.04805.

Elnagar, A., Khalifa, Y.S., Einea, A., 2018. Hotel Arabic-reviews dataset construction for sentiment analysis applications. In: Shaalan, K., Hassanien, A.E., Tolba, F. (Eds.), Intelligent Natural Language Processing : Trends and Applications. Springer International Publishing, pp. 35–52. https://doi.org/10.1007/978-3-319-67056-0_3.

Elnagar, A., Al-Debsi, R., Einea, O., 2020. Arabic text classification using deep learning models. Inf. Process. Manag. 57 (1), 102121 https://doi.org/10.1016/j.ipm.2019.102121.

Farha, I.A., Magdy, W., 2021. A comparative study of effective approaches for Arabic sentiment analysis. Inf. Process. Manag. 58 (2), 102438 https://doi.org/10.1016/j.ipm.2020.102438.

Freund, Y., Schapire, R.E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. 55 (1), 119–139. https://doi.org/10.1006/jcss.1997.1504.

Garouani, M., Kharroubi, J., 2022a. MAC : an open and free Moroccan Arabic Corpus for sentiment analysis. In: Ben Ahmed, M., Boudhir, A.A., Karas, I.R., Jain, V., Mellouli, S. (Eds.), Innovations in Smart Cities Applications, 393. Springer International Publishing, pp. 849–858. https://doi.org/10.1007/978-3-030-94191-8_68, 5.

Garouani, M., Kharroubi, J., 2022b. Towards a new lexicon-based features vector for sentiment analysis : application to Moroccan Arabic tweets. In: Maleh, Y., Alazab, M., Gherabi, N., Tawalbeh, L., Abd El-Latif, A.A. (Eds.), Advances in Information, Communication and Cybersecurity, 357. Springer International Publishing, pp. 67–76. https://doi.org/10.1007/978-3-030-91738-8_7.

Go, A., Bhayani, R., & Huang, L. (2022). Twitter Sentiment Classification Using Distant Supervision.

Habbat, N., Anoun, H., Hassouni, L., 2021. A Novel Hybrid Network for Arabic Sentiment Analysis Using Fine-Tuned AraBERT Model, 12. https://doi.org/10.15676/ijeei.2021.13.4.3.

Habbat, N., Anoun, H., Hassouni, L., 2022. Combination of GRU and CNN deep learning models for sentiment analysis on French customer reviews using XLNet model. IEEE Eng. Manag. Rev. 1–9 https://doi.org/10.1109/EMR.2022.3208818.

Hadwan, M., Al-Hagery, A., Al-Sarem, M.M., Saeed, F., 2022. Arabic sentiment analysis of users' opinions of governmental mobile applications. Comput. Mater. Continua (CMC) 72 (3), 4675–4689. https://doi.org/10.32604/cmc.2022.027311.

Hassine, M., Boussaid, L., Messaoud, H., 2016. Maghrebian dialect recognition based on support vector machines and neural network classifiers. Int. J. Speech Technol. 19 (4), 687–695. https://doi.org/10.1007/s10772-016-9360-6.

Hicham, N., Karim, S., 2022. Analysis of unsupervised machine learning techniques for an efficient customer segmentation using clustering ensemble and spectral clustering. Int. J. Adv. Comput. Sci. Appl. 13 (10) https://doi.org/10.14569/IJACSA.2022.0131016.

Hicham, N., Karim, S., Habbat, N., 2022. An efficient approach for improving customer Sentiment Analysis in the Arabic language using an Ensemble machine learning technique. In: 2022 5th International Conference on Advanced Communication Technologies and Networking (CommNet), pp. 1–6. https://doi.org/10.1109/CommNet56067.2022.9993924.

Ji Min, K., Seo Yeon, L., Won Sang, L., 2023. Discovering AI-enabled convergences based on BERT and topic network. KSII Transactions on Internet and Information Systems 17 (3). https://doi.org/10.3837/tiis.2023.03.018.

Jian, C., Gao, J., Ao, Y., 2016. A new sampling method for classifying imbalanced data based on support vector machine ensemble. Neurocomputing 193, 115–122. https://doi.org/10.1016/j.neucom.2016.02.006.

Jihad, K.H., Baker, M.R., Farhat, M., Frikha, M., 2023. Machine learning-based social media text analysis : impact of the rising fuel prices on electric vehicles. In: Abraham, A., Hong, T.-P., Kotecha, K., Ma, K., Manghirmalani Mishra, P., Gandhi, N. (Eds.), Hybrid Intelligent Systems. Springer Nature Switzerland, pp. 625–635.

Kamel/DarijaBERT · 2022 *Hugging Face*. (s. d.). Consulté 27 mai 2022, à l'adresse https://huggingface.co/Kamel/DarijaBERT.

Mann, S., Arora, J., Bhatia, M., Sharma, R., Taragi, R., 2023. Twitter sentiment analysis using enhanced BERT. In: Kulkarni, A.J., Mirjalili, S., Udgata, S.K. (Eds.), Intelligent Systems and Applications. Springer Nature Singapore, pp. 263–271.

Matthews, B.W., 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim. Biophys. Acta Protein Struct. 405 (2), 442–451. https://doi.org/10.1016/0005-2795(75)90109-9.

Mhamed, M., Sutcliffe, R., Sun, X., Feng, J., Almekhlafi, E., Retta, E.A., 2021. Improving Arabic sentiment analysis using CNN-based architectures and text preprocessing. Comput. Intell. Neurosci. 2021, 1–12. https://doi.org/10.1155/2021/5538791.

MSDA. (2022). Consulté 24 mai 2022, à l'adresse https://msda.um6p.ma/msda_datasets.

Muslim, M.A., Nikmah, T.L., Pertiwi, D.A.A., Subhan, Jumanto, Dasril, Y., Iswanto, 2023. New model combination meta-learner to improve accuracy prediction P2P lending with stacking ensemble learning. In: Intelligent Systems with Applications, 18. https://doi.org/10.1016/j.iswa.2023.200204.

Ogul, H.A., Guran, A., 2019. Imbalanced Dataset Problem in Sentiment Analysis. *2019 4th International Conference on Computer Science and Engineering*. UBMK), pp. 313–317. https://doi.org/10.1109/UBMK.2019.8907041.

Omar, N., Albared, M., Al-Shabi, A., Al-Moslmi, T., 2013. Ensemble of Classification Algorithms for Subjectivity and Sentiment Analysis of Arabic Customers ' Reviews.

Omari, M.A., 2022. OCLAR: logistic regression optimisation for Arabic customers' reviews. Int. J. Bus. Intell. Data Min. 20 (3), 251–273. https://doi.org/10.1504/IJBIDM.2022.122177.

Pota, M., Ventura, M., Catelli, R., Esposito, M., 2020. An effective BERT-based pipeline for twitter sentiment analysis : a case study in Italian. Sensors 21 (1), 133. https://doi.org/10.3390/s21010133.

Roshan, S.E., Asadi, S., 2020. Improvement of Bagging performance for classification of imbalanced datasets using evolutionary multi-objective optimization. Eng. Appl. Artif. Intell. 87, 103319 https://doi.org/10.1016/j.engappai.2019.103319.

Sagi, O., Rokach, L., 2018. Ensemble learning : a survey. WIREs Data Mining and Knowledge Discovery 8 (4), e1249. https://doi.org/10.1002/widm.1249.

Sarkar, K., 2020. A stacked ensemble approach to Bengali sentiment analysis. In: Tiwary, U.S., Chaudhury, S. (Eds.), Intelligent Human Computer Interaction. Springer International Publishing, pp. 102–111.

Sherstinsky, A., 2020. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. Phys. Nonlinear Phenom. 404, 132306 https://doi.org/10.1016/j.physd.2019.132306.

Skrobek, D., Krzywanski, J., Sosnowski, M., Kulakowska, A., Zylka, A., Grabowska, K., Ciesielska, K., Nowak, W., 2020. Prediction of sorption processes using the deep learning methods (long short-term memory). Energies 13 (24), 6601. https://doi.org/10.3390/en13246601.

Skrobek, D., Krzywanski, J., Sosnowski, M., Kulakowska, A., Zylka, A., Grabowska, K., Ciesielska, K., Nowak, W., 2022. Implementation of deep learning methods in prediction of adsorption processes. Adv. Eng. Software 173, 103190. https://doi.org/10.1016/j.advengsoft.2022.103190.

Statista The Statistics Portal. (2011). Statista. Consulté 19 mai 2022, à l'adresse https://www.statista.com/.

Tan, K.L., Lee, C.P., Anbananthen, K.S.M., Lim, K.M., 2022a. RoBERTa-LSTM : a hybrid model for sentiment analysis with transformer and recurrent neural network. IEEE Access 10, 21517–21525. https://doi.org/10.1109/ACCESS.2022.3152828.

Tan, K.L., Lee, C.P., Lim, K.M., Anbananthen, K.S.M., 2022b. Sentiment analysis with ensemble hybrid deep learning model. IEEE Access 10, 103694–103704. https://doi.org/10.1109/ACCESS.2022.3210182.

Twitter US *Airline Sentiment*. (2011). Consulté 24 janvier 2023, à l'adresse https://www.kaggle.com/datasets/crowdflower/twitter-airline-sentiment.

Williams, R.J., Zipser, D., 1989. A learning algorithm for continually running fully recurrent neural networks. Neural Comput. 1 (2), 270–280. https://doi.org/10.1162/neco.1989.1.2.270.

Xu, G., Meng, Y., Qiu, X., Yu, Z., Wu, X., 2019. Sentiment analysis of comment texts based on BiLSTM. IEEE Access 7, 51522–51532. https://doi.org/10.1109/ACCESS.2019.2909919.

Xu, Q., Tang, W., Teng, F., Peng, W., Zhang, Y., Li, W., Wen, C., Guo, J., 2019. Intelligent syndrome differentiation of traditional Chinese medicine by ann : a case study of chronic obstructive pulmonary disease. IEEE Access 7, 76167-76175. https://doi.org/10.1109/ACCESS.2019.2921318.

Zhang, A., Lipton, Z.C., Li, M., Smola, A.J., 2011. Dive into Deep Learning (s. d.).