Using AraGPT and ensemble deep learning model for sentiment analysis on Arabic imbalanced dataset

Nassera Habbat[1*], *Hicham* Nouri[2], *Houda* Anoun[1] and *Larbi* Hassouni[1]

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

Abstract. With the fast growth of mobile technology, social media has become important for people to share their thoughts and feelings. Businesses and governments can make better strategic decisions when they know what the public thinks. Because of this, sentiment analysis is an important tool for figuring out how different people's opinions are. This article presents a deep-learning ensemble model for sentiment analysis. The ensemble model proposed consists of three deep-learning models, Gated Recurrent Unit (GRU), Long Short-Term Memory (LSTM), and Bidirectional LSTM (BiLSTM), as base classifiers. AraBERT is responsible for presenting the textual input data into representative embeddings. The stacking ensemble model then captures the long-range dependencies in the embedding for a given class. As a meta-classifier, Support Vector Machine (SVM) then combines the predictions made by the stacking deep learning model. In addition, data augmentation with AraGPT was implemented to address the imbalanced dataset issues. The experimental results demonstrate that the proposed model outperforms the state-of-the-art models with an accuracy of 88.89%, 90.88%, and 88.23% on the HARD, BRAD, and Twitter datasets, respectively.

1 Introduction

In recent years, sentiment analysis has become a popular topic due to its vast array of applications. Opinion mining is another name for sentiment analysis, which uses NLP processing and deep learning techniques to systematically identify specific emotions and subjective data. Sentiment analysis examines the polarity and emotion of written texts to determine whether they are positive, neutral, or negative.

Using unbalanced data in classification will impact the learning performance of algorithms that tend to favour the majority group and result in a high misclassification rate for the minority group.

The problem mentioned above of classifying unbalanced data has attracted the attention of numerous researchers, who have proposed various solutions. The proposed methods emphasize an improved classification of a minority group. Important works proposed the methods above are as follows:

Ogul et Guran [1] compared the under-sampling and over-sampling techniques to deal three imbalanced sentiment analysis datasets (Turkish and English) used for binary classification. Consequently, it has been determined that the sample increase techniques

* Corresponding author: nassera.habbat@gmail.com

(ROS, SMOTE) boost the classification model performance values, whereas the sample reduction techniques (RUS and NM) reduce the classification performance results on the datasets using Logistic regression as a classifier. However, Albahli [2] have proposed a model for identifying authentic COVID-19-related news in Arabic Text employing sentiment-based Twitter data for Gulf countries. The suggested model for sentiment analysis employs Machine Learning and SMOTE for tackling unbalanced datasets. They obtained the best accuracy (91%) using SMOTE with Multinomial Naïve Bayes compared with baseline models.

Numerous studies have shown that ensemble learning approaches [3], [4] perform better than a single classifier when the dataset is imbalanced. Chugai et al. [5] presented a method for addressing imbalanced data classification issues by employing decision tree ensemble learning with boosting and bagging methods to construct cost-sensitive models that adjust for misclassification. Their results showed that the appropriate ensemble techniques are boosting, namely, RUSBoost, LogitBoost, TotalBoost, and AdaBoostM1, compared to the Bag model, especially, RUSBoost is the best model for classifying imbalanced data with overlapping classes and a high unbalance ratio. In addition, Tan et al. [6] presented an ensemble model that consists of three deep learning models: the combination of the Robustly optimized Bidirectional Encoder Representations from Transformers approach (RoBERTa), Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), and Bidirectional LSTM (BiLSTM), with the ensemble learning technique consisting of averaging ensemble and majority voting. Additionally, pre-trained GloVe word embeddings have been applied to the data to help even out the imbalances in the datasets. Their experimental findings demonstrated that performance is enhanced when predictions are combined in ensemble models, with both averaging and majority voting achieving a higher degree of accuracy (89.81%).

Furthermore, David et al. [7] used SMOTE to oversample minority class embeddings obtained from the BERT pretrained language model. Then, they employed the oversampled embeddings to train the Bi-LSTM classifier model to classify tweets into four classes. Their experiments demonstrate utilizing SMOTE on the top layer representations of BERT enhances the F1 score significantly more than simply adjusting the class weights.

Data augmentation is a machine learning method that augments training data with label-preserving transformations. Due to their superior performance, pre-trained language models have gained widespread acceptance in recent years. Masked language models (MLMs) like BERT and RoBERTa can forecast masked words in text according to context, allowing for text data augmentation. Abonlzio et al. [8] compared various text augmentation methods (back-translation (BT), easy data augmentation (EDA), pretrained data augmentor (PREDATOR), and BART) with latest classification algorithms (BERT, LSTM, CNN, support vector machine, GRU, and enhanced language representation with informative entities (ERNIE)). The outcomes demonstrated enhancements from the augmented dataset, particularly for smaller datasets. ERNIE and BERT performed the best with small datasets, with BT augmentation boosting the performance of BERT's classifier by 21%. In addition, the BT augmentation method's contribution with all classifiers and PREDATOR in unbalanced cases is noteworthy.

Given the importance of data augmentation on sentiment analysis tasks in case of imbalanced datasets, this paper proposes an ensemble model which integrates AraGPT as a data augmentation method, AraBERT [9] as a word embedding model, and a stacked ensemble model for the classification of short Arabic text.

The remainder of the paper's format is: The proposed ensemble model's process flow, which includes data augmentation, the word embedding model, and the ensemble method, is described in Section II. The third section describes the performance evaluation datasets and the analysis of the results. Section IV concludes the paper as a whole.

2 The proposed model

This paper presents a model for the sentiment analysis of unbalanced Arabic data sets. As shown in Figure 1, the proposed method utilizes LSTM, GRU, and BiLSTM as base classifiers, followed by a meta-classifier to aggregate the results. Using stacked ensemble learning, we can profit from each algorithm's functional and structural advantages while improving performance. In the following paragraphs, we will examine the data augmentation technique (AraGPT), the word embedding model (AraBERT), and the stacking model in greater depth.

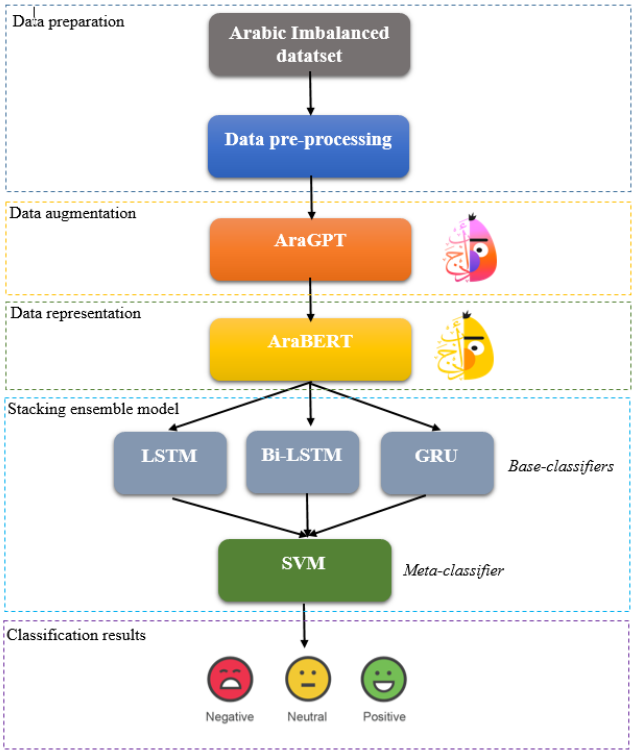

Fig. 1. The overall architecture.

2.1. AraGPT for data augmentation

GPT-2 is a sophisticated Language Model trained on 40GB of WebText and based on Transformer Architecture. Multiple decoder units are stacked on top of one another and equipped with advanced learning concepts such as Masked Self Attention, Layer Normalization, Multiple Heads, and Residual Connections among others.

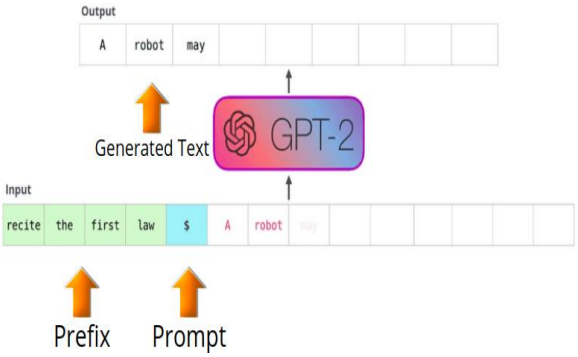

Fig. 2. Text generation of GPT-2.

GPT-2 language model attempts to optimize the ability to predict the next word in a given sequence by analyzing previous words. The illustration above (Fig. 2), taken from The depicted GPT-2 (Visualizing Transformer Language Models), provides a clear visual representation.

In our work, we used a pre-trained AraGPT-2-based model [10] to generate Arabic text from the dataset records to be modelled, resulting in a new dataset containing the generated Arabic text for the transformer (i.e., the AraGPT-2-base). We began by concatenating labels and text to create input samples, then passed to the GPT-2 model to learn the word-word and label-text dependency relations. Attaching the label to the actual sample would guide the model to control the text generation and ensure that it remains specific to a given label.

2.2. AraBERT

In this section, we examine the effect of context word embedding on sentiment analysis tasks. Specifically, we employed the AraBERT model [11], an Arabic pre-training BERT transformer model that is a deep, unsupervised, bidirectional language representation that can generate word embeddings to represent the semantics of words in their context. It is pre-trained with datasets from Arabic news websites for articles; approximately 1 billion tokens in 3.5 million articles from the Open Source International Arabic News (OSLAN) Corpus and 1.5 billion words in 5 million articles from 10 major news sources in eight countries.

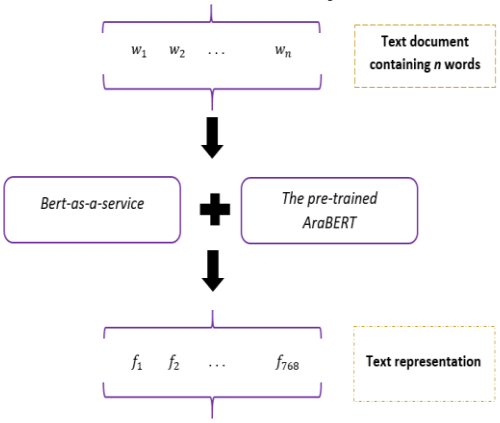

Fig. 3. Extraction of features from the pre-trained AraBERT.

Figure 3 depicts the best-as-a-service technology, which activates one or more layers without adjusting the AraBERT parameters. It computes the average pool of all tokens'

second-to-last hidden layer. The output representation becomes the input for the stacking ensemble model we will discuss in the following section.

2.3. Stacking ensemble model

Ensemble techniques employ multiple learning algorithms to generate a single optimal predictive model. The model's performance is superior to that of the learners used alone. The primary classifications for ensemble techniques are Bagging, Boosting, and Stacking.

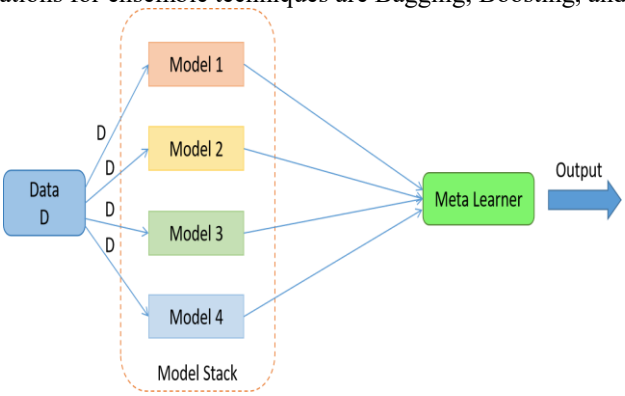

Fig. 4. The stacking algorithm.

Stacking frequently takes into account heterogeneous learners, trains them in parallel, and merges them by training a meta-learner to generate a prediction based on the predictions of the various learners. A meta learner receives as features the predictions and the target as the ground truth values in data D (Fig. 4). It then tries to figure out how to best combine the input predictions to generate a more improved prediction.
In our work, we used three deep-learning algorithms as base classifiers and the SVM algorithm as meta-classifier.

3 Experiments and results analysis

This section discusses the Arabic datasets used in this study, followed by the experimental outcomes.

3.1 Datasets

To assess the efficacy of our proposed method, we apply the model to three Arabic imbalanced datasets, as summarized in Table 1.

- Hotel Arabic-Reviews Dataset (HARD) [12]: This dataset contains 93,700 Arabic-language hotel reviews. During June and July 2016, the hotel reviews were collected from the Booking.com website. The comments are written in both Modern Standard Arabic (MSA) and dialectal Arabic. We utilized the unbalanced dataset, which consists of 373,750 reviews, in our research. This is a clean dataset containing all reviews.
- Hotel Arabic-Reviews Dataset (BARD) [13]: This dataset contains 510,600 Arabic-language book reviews. The reviews were gathered from the GoodReads.com website between June and July 2016. This work is an expansion of the initial large-scale Arabic dataset, LABR, which consists of approximately 63,000 Arabic Book Reviews collected from GoodReads.com. The reviews are predominantly written in MSA, but there are also

reviews written in dialectal Arabic. In our study, we utilized an unbalanced dataset containing over 510K reviews. This is a preprocessed dataset containing all reviews.

- Twitter dataset (TD) [2]: The dataset included Arabic tweets posted intermittently throughout the COVID-19 pandemic. This study's data were limited to the Gulf states, namely Qatar, Oman, Bahrain, United Arab Emirates (UAE), and the Saudi Arabia. Beginning in March of 2020 and ending in April 2020, the dataset was crawled from Twitter. 60,000 tweets were gathered using the keyword search coronavirus, corona, covid19, sarscov2, and COVID

Table 1. The used imbalanced datasets.

Dataset	Positive	Negative	Neutral	Total
HARD	18 300 60%	5 490 18%	6 710 22%	30 500
BRAD	46 900 67%	9 100 13%	14 000 20%	70 000
TD	6 690 30%	4 014 18%	11 596 52%	22 300

Note that we did not use the entire described datasets. The used datasets are summarized in Table 1.

3.2 Results analysis

To demonstrate the effectiveness of our proposed model, we applied it to the aforementioned datasets and compared stacking model with single classifiers, including GRU, LSTM, and BiLSTM. Regarding the phase of data augmentation (DA), we compared AraGPT to SMOTE and RUS. We utilized 30 epochs and a batch size of 64 for training.
As performance metrics, we employed accuracy, F1 score, and MCC [14] [15].

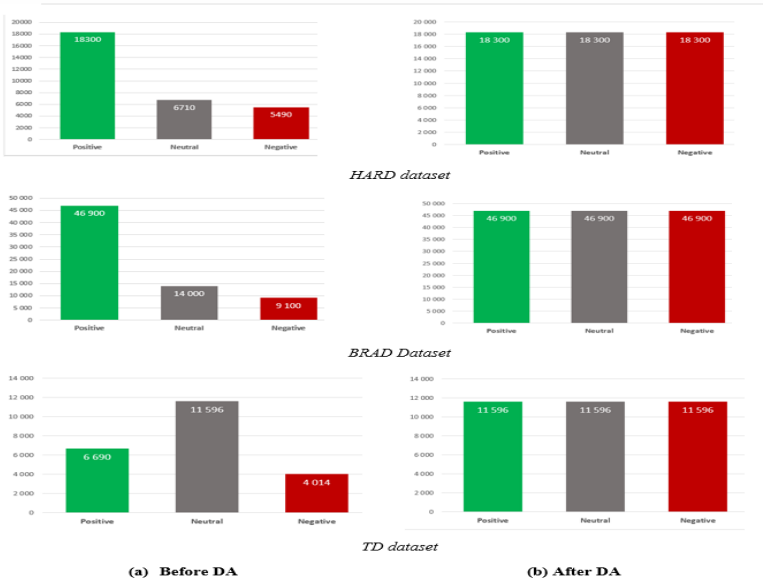

Fig. 5. The datasets distribution (a) Before DA and (b) After DA.

Figure 5 shows the distribution of class samples in the datasets before and after DA. All datasets are split into training sets of 80% and testing sets of 20%.

Table 2. Results of comparative experiments on the HARD dataset.

Model	DA technique	Accuracy	F1-score	MCC
LSTM	**SMOTE**	77.20	77	77.11
	RUS	73.90	73.20	73
	AraGPT	85.25	84.97	84.01
Bi-LSTM	**SMOTE**	78.15	78	78
	RUS	74.02	73.84	73
	AraGPT	86.46	86.12	86
GRU	**SMOTE**	79.62	79.50	78
	RUS	75.64	75.60	75.12
	AraGPT	87.84	87.82	87.50
Stacked model	**SMOTE**	80.10	80	80.03
	RUS	79.30	79.22	79.14
	AraGPT	88.89	88.80	88

Table 3. Results of comparative experiments on the BRAD dataset.

Model	DA technique	Accuracy	F1-score	MCC
LSTM	**SMOTE**	78.89	78.10	78
	RUS	74.03	74	74.26
	AraGPT	86.78	86.81	86
Bi-LSTM	**SMOTE**	79.40	79	79.02
	RUS	74.50	74.87	74
	AraGPT	87.50	87.10	87
GRU	**SMOTE**	80.61	80.46	80
	RUS	76.75	76.62	76.16
	AraGPT	88.80	88.78	88
Stacked model	**SMOTE**	80.16	80	80
	RUS	79.46	79.35	79.19
	AraGPT	**90.88**	**89.79**	**89**

Table 4. Results of comparative experiments on the TD dataset.

Model	DA technique	Accuracy	F1-score	MCC
LSTM	**SMOTE**	73	73.01	73
	RUS	71.68	71.24	71.22
	AraGPT	83.78	83.11	83
Bi-LSTM	**SMOTE**	74.15	74.09	74
	RUS	73.45	73	73.11
	AraGPT	85.92	85	85.50
GRU	**SMOTE**	77.61	77.45	77.10
	RUS	74.79	74.65	74
	AraGPT	86.80	86.84	86
Stacked model	**SMOTE**	79.11	79	79.23
	RUS	78.77	78.67	78
	AraGPT	88.23	88	88.02

On the one hand, utilizing AraGPT as an enhancement method improves the model's performance and generates extremely realistic samples. In addition, it has been discovered that the sample increase method (SMOTE) increases classifier performance values relative

to the sample reduction method (RUS) and decreases data set performance values. The results are elaborately explained.

On the other hand, merging the base deep learning models in a stacked technique improves classification accuracy relative to each classifier. It implies that merging models with diverse functional and structural qualities can be advantageous. Overall, using AraGPT as a data augmentation technique, AraBERT as a word embedding model, and the stacking model for Arabic sentiment analysis improves accuracy, F1-score, and MCC by 90.88%, 89.79%, and 89%, respectively.

4 Conclusion and future directions

Even the most advanced, state-of-the-art language models encounter significant difficulties due to imbalanced datasets. In general, we address this issue with sampling techniques such as under-sampling, over-sampling, or transformer-based models. In this work, we employ AraGPT for data augmentation, AraBERT for text representation, and a stacking deep learning model for classification. Here are some empirical results:

1. Over-sampling methods like SMOTE outperform under-sampling methods like RUS.
2. Using AraGPT generates very realistic samples and improves the classification task.
3. Using the stacking model increases the performance of the classification task.

In future directions, we aim to use GPT for data augmentation of a small dataset and test its efficiency using different combinations for classification.

References

1. H. A. Ogul et A. Guran, « Imbalanced Dataset Problem in Sentiment Analysis », in *2019 4th International Conference on Computer Science and Engineering (UBMK)*, Samsun, Turkey, sept. 2019, p. 313-317. doi: 10.1109/UBMK.2019.8907041.
2. S. Albahli, « Twitter sentiment analysis: An Arabic text mining approach based on COVID-19 », *Front. Public Health*, vol. 10, p. 966779, oct. 2022, doi: 10.3389/fpubh.2022.966779.
3. N. Hicham, S. Karim, et N. Habbat, « 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)*, 2022, p. 1-6. doi: 10.1109/CommNet56067.2022.9993924.
4. N. Hicham et S. Karim, « Analysis of Unsupervised Machine Learning Techniques for an Efficient Customer Segmentation using Clustering Ensemble and Spectral Clustering », *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, n° 10, 2022, doi: 10.14569/IJACSA.2022.0131016.
5. P. Chujai, K. Chomboon, P. Teerarassamee, N. Kerdprasop, et K. Kerdprasop, « Ensemble Learning For Imbalanced Data Classification Problem », in *The Proceedings of the 2nd International Conference on Industrial Application Engineering 2015*, 2015, p. 449-456. doi: 10.12792/iciae2015.079.
6. K. L. Tan, C. P. Lee, K. M. Lim, et K. S. M. Anbananthen, « Sentiment Analysis With Ensemble Hybrid Deep Learning Model », *IEEE Access*, vol. 10, p. 103694-103704, 2022, doi: 10.1109/ACCESS.2022.3210182.
7. J. David, J. Cui, et F. Rahimi, « CLASSIFICATION OF IMBALANCED DATASET USING BERT EMBEDDINGS », 2020.

8. H. Q. Abonizio, E. C. Paraiso, et S. Barbon, « Toward Text Data Augmentation for Sentiment Analysis », *IEEE Trans. Artif. Intell.*, vol. 3, nº 5, p. 657-668, oct. 2022, doi: 10.1109/TAI.2021.3114390.
9. N. Habbat, H. Anoun, et L. Hassouni, « A Novel Hybrid Network for Arabic Sentiment Analysis using fine-tuned AraBERT model », p. 12, 2021, doi: 10.15676/ijeei.2021.13.4.3.
10. W. Antoun, F. Baly, et H. Hajj, « AraGPT2: Pre-Trained Transformer for Arabic Language Generation », p. 12.
11. W. Antoun, F. Baly, et H. Hajj, « AraBERT: Transformer-based Model for Arabic Language Understanding », p. 7.
12. A. Elnagar, Y. S. Khalifa, et A. Einea, « Hotel Arabic-Reviews Dataset Construction for Sentiment Analysis Applications », in *Intelligent Natural Language Processing: Trends and Applications*, K. Shaalan, A. E. Hassanien, et F. Tolba, Éd. Cham: Springer International Publishing, 2018, p. 35-52. doi: 10.1007/978-3-319-67056-0_3.
13. M. Aly et A. Atiya, « LABR: A Large Scale Arabic Book Reviews Dataset », 2013, doi: 10.13140/2.1.3960.5761.
14. B. W. Matthews, « Comparison of the predicted and observed secondary structure of T4 phage lysozyme », *Biochim. Biophys. Acta BBA - Protein Struct.*, vol. 405, nº 2, p. 442-451, 1975, doi: https://doi.org/10.1016/0005-2795(75)90109-9.
15. S. Boughorbel, F. Jarray, et M. El-Anbari, « Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric », *PLOS ONE*, vol. 12, nº 6, p. 1-17, juin 2017, doi: 10.1371/journal.pone.0177678.