The Scientific Temper (2024) Vol. 15 (4): 3099-3104
E-ISSN: 2231-6396, ISSN: 0976-8653
Doi: 10.58414/SCIENTIFICTEMPER.2024.15.4.17
https://scientifictemper.com/

RESEARCH ARTICLE

EMSMOTE: Ensemble multiclass synthetic minority oversampling technique to improve accuracy of multilingual sentiment analysis on imbalance data

Ayesha Shakith*, L. Arockiam

Abstract

Natural language processing (NLP) tasks, such as multilingual sentiment analysis, are inherently challenging, especially when dealing with unbalanced data. A dataset is considered imbalanced when one class significantly dominates the others, creating an unbalanced distribution. In many domains, the minority class holds crucial information, presenting unique challenges. This research addresses these challenges using an ensemble-based oversampling technique, the ensemble multiclass synthetic minority oversampling technique (EMSMOTE). By leveraging SMOTE, EMSMOTE generates multiple synthetic datasets to train various classifiers. The proposed model, when combined with an ensemble random forest classifier, attained an impressive accuracy of 90.73%. This ensemble approach not only mitigates the effects of noisy synthetic samples introduced by SMOTE but also showcases significant enhancement in the overall performance in tackling class imbalances.

Keywords: Sentiment analysis, Natural language processing, Multilingual dataset, Imbalance classification, SMOTE.

Introduction

Sentiment analysis is the process of figuring out the sentiment or emotion portrayed in a text. It has grown in significance in several industries, including marketing research, customer feedback analysis, and social media monitoring. Analyzing sentiment is a crucial method for determining how disparate people's viewpoints are [N. Habbat *et al.*, 2023]. Sentiment analysis, a task commonly studied in monolingual settings, has been extended to the multilingual framework to capture sentiment expressed in different languages. The task of multilingual sentiment analysis is particularly challenging when dealing with imbalanced data. Imbalanced data refers to the situation where there is a significant disparity in the number of samples available for different sentiment classes. This can lead to biased and inaccurate sentiment analysis results. To address this issue, various approaches have been proposed in NLP. It aims to effectively handle the imbalance in multilingual sentiment analysis datasets and improve the overall accuracy of sentiment classification. One approach is the use of data resampling techniques, such as oversampling or undersampling, to balance the distribution of sentiment classes in the dataset. Dealing with imbalanced data, or dramatically different numbers of samples in each class, is a significant challenge in sentiment analysis. To overcome the problem of imbalanced data in sentiment analysis, this research focuses on the application of ensemble learning approaches. Then, numerous classifiers are combined, as opposed to utilizing a single classifier, and the predictive accuracy is improved (an approach known as ensemble learning). When it comes to improving the classification performance of unbalanced data, ensemble learning approaches outperform conventional data sampling strategies. SMOTE is an oversampling technique that interpolates between current minority class samples to create artificial examples in the minority class. By using this technique, the performance of classifiers trained on unbalanced data is enhanced and the distribution of data is helped to balance.

Department of Computer Science, St. Joseph's College (Autonomous), Affiliated to Bharathidasan University, Trichy, India.

***Corresponding Author:** Ayesha Shakith, Department of Computer Science, St. Joseph's College (Autonomous), Affiliated to Bharathidasan University, Trichy, India, E-Mail: Ayeshasm1412@gmail.com

How to cite this article: Shakith, A., Arockiam, L. (2024). EMSMOTE: Ensemble multiclass synthetic minority oversampling technique to improve accuracy of multilingual sentiment analysis on imbalance data. The Scientific Temper, **15**(4):3099-3104.

Doi: 10.58414/SCIENTIFICTEMPER.2024.15.4.17

Source of support: Nil

Conflict of interest: None.

Received: 19/10/2024 **Accepted:** 22/11/2024 **Published** : 20/12/2024

Related Works

One notable strategy involves a hybrid approach that combines SVM with PSO and numerous oversampling methods to address the issue of imbalanced data in customer evaluations [R.Obiedat *et al.*, 2022]. Another approach proposes a hybrid deep learning technique, RoBERTa-LSTM, which leverages the strengths of both long short-term memory (LSTM) networks and the robustly tuned BERT methodology to enhance sentiment analysis performance [Kian Long Tan *et al.*, 2022].

To further improve the handling of imbalanced datasets, a model integrating traditional machine learning methods with the synthetic minority over-sampling technique (SMOTE) has been suggested. This combination proves effective in dealing with class imbalance in sentiment analysis tasks [Salah Albahli *et al.*, 2022]. Similarly, an advanced method called CDSMOTE, which combines SMOTE with class decomposition, has been proposed to enhance the effectiveness of sentiment analysis on textual data, particularly when dealing with class imbalances [Carlos Francisco].

In the context of performance evaluation, it has been observed that selecting the optimal ratio, such as the f-score, plays a crucial role in achieving the best results. The process involves identifying the class by majority vote and subsequently removing misclassified points from the training set to refine the model [Zeeshan Ali Sayyed, 2022]. Ensemble learning has also gained traction in sentiment analysis, with a variety of techniques being explored across different datasets. For instance, a neural network-based ensemble learning approach has been shown to achieve state-of-the-art accuracy by merging language and auditory representations, particularly in multimodal sentiment analysis [Hossain Gomeshi *et al.*, 2023].

In Urdu sentiment analysis, a meta-learning ensemble technique has been developed that combines deep learning and basic machine learning models. This hybrid approach significantly improves classification performance compared to baseline deep models [Kanwal Ahmed *et al.*, 2023]. Another effective ensemble method utilizes LSTM networks as base learners and decision trees as meta-learners in a stacking ensemble model. This model outperforms other ensemble learning approaches in classifying translated text [Thuraya *et al.*, 2022]. Additionally, an improved heterogeneous stacking ensemble model has been proposed for Arabic sentiment analysis. This model integrates pre-trained deep learning models with meta-learners, thereby enhancing the overall performance of sentiment analysis [Hager Saleh, 2022].

Finally, a distance-based method has been introduced to determine the optimal number of weak classifiers required for training ensemble learning models, which plays a vital part in improving the accuracy and robustness of sentiment analysis systems [Can Ozbey, 2021].

Research Methodology

Most methods used in sentiment analysis perform optimally when fed uniformly distributed datasets that are labeled, although this is not always available when doing sentiment analysis. Different combinations of sampling techniques were then used to reduce the class imbalance and the results of these are shown and discussed [Zeeshan Ali Sayyed, 2021]. By using improved SMOTE, a balanced dataset can be achieved. Performance evaluation can be done by using the following machine learning algorithms: Random forest, support vector machine, Naïve Bayes, BiLSTM-CNN, Logistic Regression, XgBoost, BERT- multilingual, KNN as mentioned in Figure 1. The tools used for this research are as follows, The programming language used are Python 3.9. Libraries used are: Pandas, Numpy, Scipy, matplotlib, spacy, sklearn and imblearn. This study, it experiments with the issue of detecting in multilingual datasets pleasant, negative, mixed, unknown, and not-Tamil sentiments (Figures 2-4).

The Figure 2 represents the training dataset. This dataset contains five class labels. The positive labels have 67% and it is majority class-label. The mixed feelings, negative, not-Tamil, and unknown, are 12, 13, 3 and 5%, respectively, and those are the least minority classes.

Since they are less than 10% of the majority class positive, it has an imbalanced training data set that will affect the accuracy of the predictions. Two major approaches for imbalance classification are as follows: Under-sampling and over-sampling.

EMSMOTE

Ensemble multiclass synthetic minority over-sampling technique (EMSMOTE) is an extension of the SMOTE algorithm that handles multiclass imbalance. It generates synthetic samples for each minority class by using K-NN to consider its nearby minority classes in the feature space. It also generates synthetic samples for each minority class by clustering (k-means) its nearby minority classes in the feature space. The high-level diagram of the proposed work is given in the Figure 5.

The steps of EMSMOTE are as follows:

- Load the dataset
  - -Pre-processing
  - -Word embedding
  - -Feature extraction
  - -Calculate the majority and minority classes.
  - -Find the minority classes that require oversampling (Here threshold is below 10% higher than the majority class).
- Create ensemble classifiers
  - -Bagging
  - -Boosting (this work aims to develop an ensemble using boosting)
  - -Decision tree
  - -Stacking

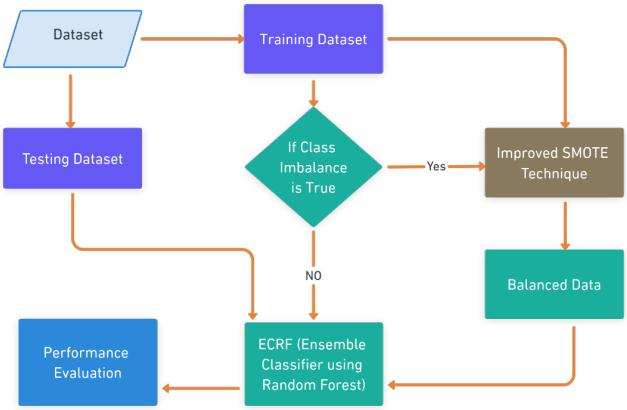

Figure 1: Methodology diagram

- Looping the oversampling using EMSMOTE.
- For each iteration generated dataset from step 3, train the ensemble classifiers and make predictions.
- Performance evaluation

ECRF

To increase the classification task's accuracy and robustness, the ensemble random forest (ECRF) algorithm combines the predictions of several decision trees. By randomly selecting features and creating bootstrap samples, it introduces diversity among the decision trees, reducing overfitting and enhancing generalization. The algorithm's strength lies in the combination of individual decision trees' predictions through majority voting, which reduces the impact of outliers or noise and produces more reliable predictions.

Baseline Models: SVM, NB, DT, RF.

Algorithm-1: Ensemble SMOTE (EMSMOTE) function

Input: Imbalance dataset

Output: Balanced dataset

- Load D
- D′ = [ ]
- For S in D:
  - Tokenize each word in a sentence
  - Remove punctuation, stop words, URL links, emoticons and special characters.
  - Apply lemmatization
  - Apply stemming
  - Append S to D′
- Load D′
- Apply cross-lingual word embeddings
  - E x E $\in \mathbb{R}^{|V_E| \times d_E}$ represent the matrix of English word embeddings.
  - E X T $\in \mathbb{R}^{|V_T| \times d_T}$ represents the matrix of Tamil word embeddings.
  - Apply max pooling
- $D_{syn} = \{\}$
- If $C_{maj} > C_{min}$:
  - Calculate C_{min}
- for min_set in C_{min}:

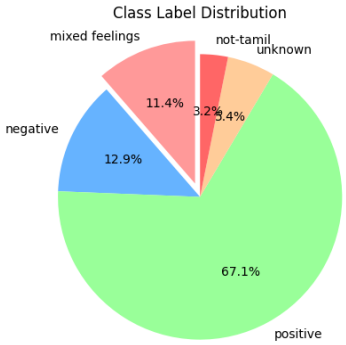

Figure 2: Training dataset with multiple classifiers

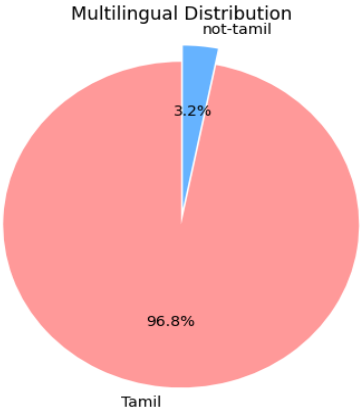

Figure 3: Multilingual distribution

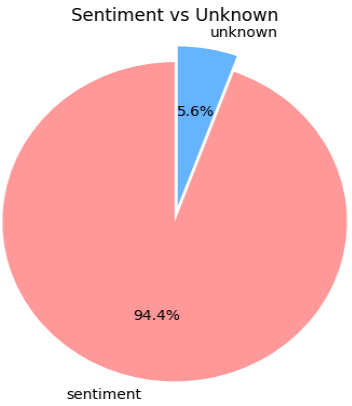

Figure 4: Sentiment vs unknown

- For instance in C_{min}:
  - Identify the k nearest neighbours of each minority class sample within the same class
  - If random KNN is true:
- $Diff_{FV =}$ calculate the difference between the feature vector of the current instance with random KNN value
- Res = $Diff_{FV}$ * Rand (0,1)
- Add res to the feature vector of the instance
  - Store the instance in D_{syn}

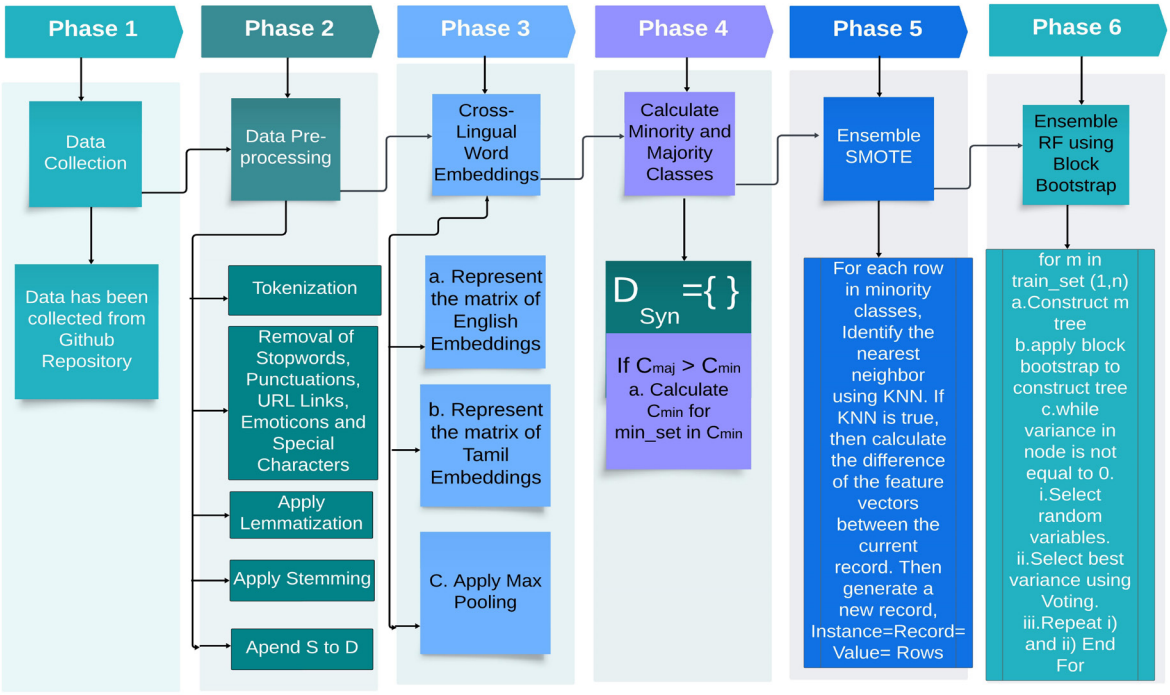

Figure 5: Overall architecture of EMSMOTE

- Load the synthetic samples into ensemble classifiers ()
- Calculate the confusion matrix
- Identify the best model

Notations used

- D → dataset
- S → sentence
- D′ → pre-processed data
- E x E → matrix of English word embeddings.
- E X T → matrix of Tamil word embeddings.
- V_E → vocabulary of English words
- V_T → vocabulary of Tamil words
- d_E → dimensionality of English word embeddings
- d_T → dimensionality of Tamil word embeddings
- C_{maj} → majority class
- C_{min} → minority class
- D_{syn} → generated synthetic samples
- Min_set → least minority class dataset

Algorithm-2: Ensemble Classifier using Random Forest
Input: Synthetic dataset
Output: Predictions

- load dataset
- Split the dataset into training and test data
- train_set = $x_i \in X$, i=1,2,3,..,n with responses $y_i \in Y$, i=1,2,3,..,n
- for m in train_set(1,n):
  - construct m^{th} tree
  - Apply block bootstrap to construct a tree
  - while the variance in node ≠ 0:
  - Select random variables
  - Select the best variance using voting
  - Repeat steps i and ii
- end for

Baseline Models: SVM, NB, DT, RF.

Results and Discussion

Comparison of ECRF with RF, DT, NB, and SVM.

In the first experiment, the performance of an ECRF algorithm is compared with four commonly used classification algorithms: RF, DT, NB, and SVM. The goal is to assess the effectiveness of the improved random forest algorithm compared to these established methods. To compare these algorithms, the above-mentioned dataset is used for training and evaluation.

The choice of the most suitable algorithm would depend on the specific requirements and trade-offs in the given problem domain. In terms of overall accuracy, the ECRF improves by 4% when compared to SVM and 1% when compared to the RF algorithm. The most noticeable feature of Table 1, is that the recall is greater than precision. The ratio of accurately predicted positive cases to the total number of actual positive instances is called recall, which is also referred to as sensitivity or true positive rate. It assesses how well the model can recognize every positive case. Contrarily, precision is defined as the ratio of accurately anticipated positive instances to all expected positive instances. It assesses how well the model can distinguish positive examples from all the occurrences it anticipated to be positive. In an imbalanced dataset scenario, where the majority class dominates the dataset, a model can achieve

Table 1: Comparative analysis of proposed ECRF with baseline models

	Precision	*Recall*	*F-score*	*Accuracy*
SVM	81.87	86.96	84.338	82.45
NB	82.83	87.9	85.289	83.4
DT	83.76	88.81	86.21	84.32
RF	84.72	89.75	87.16	85.27
ECRF	85.69	90.7	88.12	86.23

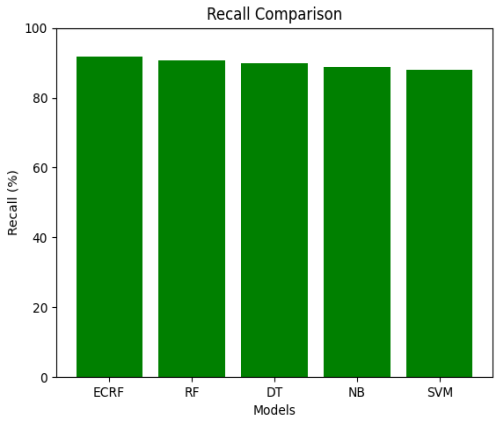

Figure 6: Recall comparison

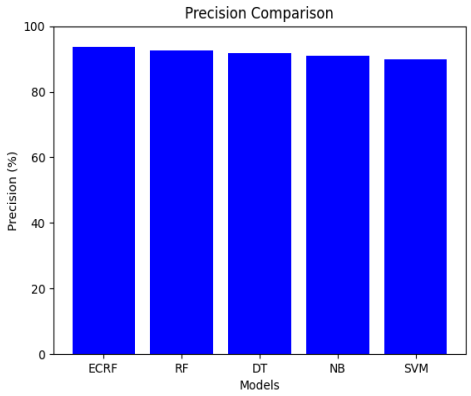

Figure 7: Precision comparison

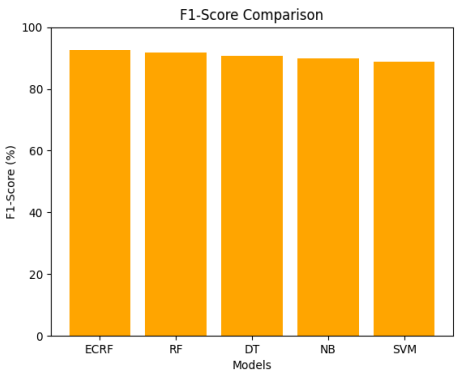

Figure 8: F1- score comparison

Table 2: Comparative analysis of proposed EMSMOTE with ECRF and baseline models

	Precision	*Recall*	*F-score*	*Accuracy*
EMSMOTE-SVM	89.87	87.96	88.9	87.01
EMSMOTE-NB	90.93	88.899	89.85	87.96
EMSMOTE-DT	91.76	89.809	90.7745	88.884
EMSMOTE-RF	92.72	90.749	91.72	89.83
EMSMOTE-ECRF	93.69	91.699	92.68	90.73

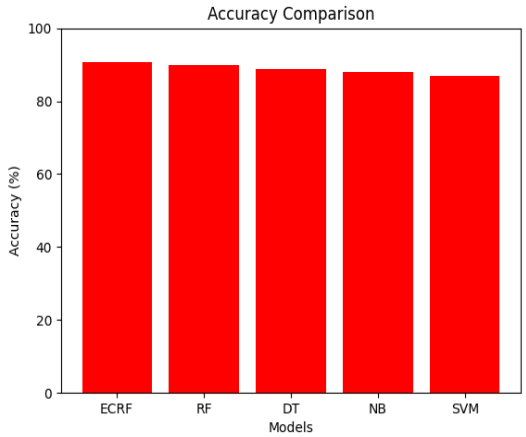

Figure 9: Accuracy comparison

a high recall by predicting the majority class as positive for most instances, thus capturing a significant portion of the actual positive instances. However, this approach may result in a lower precision because the model is more likely to misclassify negative instances as positive, given their abundance in the dataset.

Comparison after Applying the Ensemble SMOTE Technique

In the second experiment it conducts a comparison after applying the proposed ensemble SMOTE to the dataset. Ensemble SMOTE is a data augmentation technique widely used to address class imbalance in classification problems.

To balance the dataset and enhance classifier performance, it creates synthetic samples for the minority class. The enhanced dataset is utilized to assess the same set of algorithms, and the evaluation criteria are employed to gauge each algorithm's performance. When it comes to overall accuracy, the ECRF outperforms SVM by 3% and the RF algorithm by 1.1%. The fact that precision exceeds recall is the most noteworthy feature of Table 2. This is because there are enough of both positive and other classes. Both accuracy and recall are significant evaluation criteria in a well-balanced dataset. Figures 6-9 represent the comparative results of recall, precision, f1-score, and accuracy, respectively.

Conclusion

Using the proposed Ensemble SMOTE technique enhances the performance of all the ML models. However, even after applying the technique, the value of the ECRF is higher than those of the other algorithms. It proves its advantage in dealing with class imbalance and introduces the potential of the algorithm and the data augmentation method in enhancing the prediction result. Based on the results it can be concluded that all the algorithms benefit when using the ensemble SMOTE approach. This demonstrates how well it deals with class imbalance and that both the algorithm and data augmentation strategies yield improved performance in predictions.

Acknowledgment

The UGC provided funds for this study as part of the "Savitribai Jyotirao Phule Single Girl Child Fellowship (SJSGC) 2022-23" initiative. The author thanks the UGC for its financial support and the significant role it played in the successful completion of this work. Opinions, research, and ideas made by the author are solely those of the author and may not necessarily represent those of the UGC.

References

Ahmed, K., Nadeem, M. I., Li, D., Zheng, Z., Al-Kahtani, N., Alkahtani, H. K., Mostafa, S. M., & Mamyrbayev, O. (2023). Contextually enriched meta-learning ensemble model for Urdu sentiment analysis. *Symmetry, 15*(3), 645. https://doi.org/10.3390/sym15030645

Ali Sayyed, Z. (2021). Study of sampling methods in sentiment analysis of imbalanced data. *arXiv: Computation and Language*. https://arxiv.org/abs/2106.06673v1

Albahli, S. (2022). Twitter sentiment analysis: An Arabic text mining approach based on COVID-19. *Frontiers in Public Health, 10*. https://doi.org/10.3389/fpubh.2022.944069

Alshamsi, A., Bayari, R., & Salloum, S. A. (2020). Sentiment analysis in English texts. *Advances in Science, Technology and Engineering Systems Journal (ASTES Journal), 5*(6), 1683-1689. https://doi.org/10.25046/aj050621

Bhattacharjee, M., Ghosh, K., Banerjee, A., & Chatterjee, S. (2021). Multilabel sentiment prediction by addressing imbalanced class problem using oversampling. *Journal of Computing and Information Science in Engineering, 21*(1), 239-249. https://doi.org/10.1115/1.4049441

Faris, H., Harfoushi, O., Al-Qaisi, L., Al-Zoubi, A., Ala', M., Obiedat, R., & Qaddoura, R. (2022). Sentiment analysis of customers' reviews using a hybrid evolutionary SVM-based approach in an imbalanced data distribution. *IEEE Access, 10*, 22260-22273. https://doi.org/10.1109/ACCESS.2022.3153948

George, S., & Srividhya, V. (2022). Performance evaluation of sentiment analysis on balanced and imbalanced dataset using ensemble approach. *Indian Journal of Science and Technology, 15*(17), 790-797. https://doi.org/10.17485/IJST/v15i17.1890

Ghomeshi, H., & Vakaj, E. (2023). An ensemble-learning-based technique for bimodal sentiment analysis. *Big Data and Cognitive Computing, 7*(2), 85. https://doi.org/10.3390/bdcc7020085

Habbat, N., Nouri, H., Anoun, H., & Hassouni, L. (2023). Using AraGPT and ensemble deep learning model for sentiment analysis on Arabic imbalanced dataset. *ITM Web of Conferences, 52*, 02008. https://doi.org/10.1051/itmconf/20235202008

Hanhach, H., & Benkhalifa, M. (2019). An enhanced MNB based model for explicit and hidden sentiment classification in imbalanced datasets. *International Journal of Intelligent Engineering and Systems, 12*(5), 74-84. https://doi.org/10.22266/ijies2019.1031.09

Moreno-García, C. F., Jayne, C., & Elyan, E. (2021). Class-decomposition and augmentation for imbalanced data sentiment analysis. In *Proceedings of 2021 International Joint Conference on Neural Networks (IJCNN 2021)* (pp. 18-22). IEEE. https://doi.org/10.1109/IJCNN52387.2021.9533518

Omran, T. M., Sharef, B. T., Grosan, C., & Li, Y. (2022). Ensemble learning for sentiment analysis of translation-based textual data. *Data, 8*(4), 68. https://doi.org/10.3390/data8040068

Omran, T. M., Sharef, B. T., Grosan, C., & Li, Y. (2023). Sentiment analysis of multilingual dataset of Bahraini dialects, Arabic, and English. *Data, 8*(4), 68. https://doi.org/10.3390/data8040068

Ozbey, C., Dilekoglu, B., & Aciksoz, S. (2021). The impact of ensemble learning in sentiment analysis under domain shift. *Journal of Data Science and Analytics, 15*(4), 1-11. https://doi.org/10.1007/s41060-021-00291-1

Pu, X., Yan, G., Yu, C., Mi, X., & Yu, C. (2021). Sentiment analysis of online course evaluation based on a new ensemble deep learning model: Evidence from Chinese. *Applied Sciences, 11*(23), 11313. https://doi.org/10.3390/app112311313

Saleh, H., Mostafa, S., Alharbi, A. I., El-Sappagh, S., & Alkhalifah, T. (2022). Heterogeneous ensemble deep learning model for enhanced Arabic sentiment analysis. *Sensors, 22*(10), 3707. https://doi.org/10.3390/s22103707

Shaverizade, A., & Mohammadi, A. (2021). Ensemble deep learning for aspect-based sentiment analysis. *International Journal of Nonlinear Analysis and Applications, 12*, 29-38. https://doi.org/10.22075/ijnaa.2021.4854

Srividhya, V., & George, S. (2022). Performance evaluation of sentiment analysis on balanced and imbalanced dataset using ensemble approach. *Indian Journal of Science and Technology, 15*(17), 790-797. https://doi.org/10.17485/IJST/v15i17.1890

Tan, K. L., Lee, C. P., Sonai Muthu Anbananthen, K., & Lim, K. M. (2022). RoBERTa-LSTM: A hybrid model for sentiment analysis with transformer and recurrent neural network. *IEEE Access*. https://doi.org/10.1109/ACCESS.2022.31528