applied system innovation

MDPI

*Article*

# A Hybrid SBERT–WGAN Framework with Ensemble Learning for Sentiment Analysis in Imbalanced Datasets

**Hamza Jakha [1,*], Sanae Tbaikhi [1], Souad El Houssaini [1], Mohammed-Alamine El Houssaini [2] and Souad Ajjaj [3]**

[1] ELIITS Laboratory, Department of Computer Science, Faculty of Sciences, Chouaib Doukkali University, El Jadida 24000, Morocco; tbaikhi.sanae@ucd.ac.ma (S.T.); elhoussaini.s@ucd.ac.ma (S.E.H.)
[2] Higher School of Education and Training, Chouaib Doukkali University, El Jadida 24000, Morocco; elhoussaini.m@ucd.ac.ma
[3] Department of Computer Science, Faculty of Sciences, Ibn Tofail University, Kenitra 14000, Morocco; souad.ajjaj@uit.ac.ma
* Correspondence: jakha.h@ucd.ac.ma

**Abstract**

Sentiment analysis has become increasingly important across various domains, particularly in business intelligence, where it is crucial for improving the performance of companies by identifying the sentiments and emotions expressed in customer feedback on products and services. Despite its growing relevance, sentiment analysis still faces several challenges, including class imbalance in datasets, limitations in feature extraction techniques, and the selection of appropriate classification models. Effectively addressing these challenges requires the integration of robust representation methods, reliable data balancing strategies, and efficient classification frameworks. In this study, we propose a novel sentiment analysis approach that combines SBERT for contextual feature extraction, WGAN-based synthetic data generation for addressing class imbalance, and a soft voting ensemble classifier for improved prediction. The proposed approach is evaluated on five datasets, including two English datasets and three Arabic datasets, in order to assess its performance in a multilingual setting. We compare the effectiveness of the proposed model with several baseline machine learning classifiers, as well as with commonly used data balancing techniques such as the synthetic minority over-sampling technique (SMOTE) and adaptive synthetic (ADASYN). The evaluation is conducted using multiple performance metrics, including accuracy, precision, recall, F1-score, MCC, ROC–AUC and training and inference time, along with different validation strategies including fixed train–test splits and k-fold cross-validation. The experimental results demonstrate the effectiveness and stability of the proposed approach. In particular, they highlight the importance of capturing sentence-level contextual representations and generating realistic synthetic samples to address class imbalance.

**Keywords:** sentiment analysis; business intelligence; sentence BERT; generative adversarial network; ensemble learning; imbalanced datasets



Academic Editor: María Dolores Ruiz

Received: 2 April 2026
Revised: 6 May 2026
Accepted: 14 May 2026
Published: 19 May 2026



## 1. Introduction

In recent years, the rapid growth of social networks, online platforms, and discussion forums has generated a massive volume of user-generated data. Analyzing this data has become essential for understanding consumer behavior and extracting valuable insights. This is particularly relevant in the field of business intelligence, where customer feedback plays a crucial role in improving product and service quality, as well as in guiding marketing

strategies to better meet customer expectations. This evolution has led to the emergence of sentiment analysis, a key application of natural language processing (NLP) [1] within the broader domain of artificial intelligence (AI). Sentiment analysis, also called opinion mining [2], focuses on identifying and interpreting opinions and emotions expressed in many forms of data. In the context of business intelligence, it is widely applied to analyze customer reviews and feedback regarding services such as hotels, restaurants, and other consumer-oriented platforms.

The standard pipeline of sentiment analysis in textual data generally follows a common sequence of steps, regardless of the application domain, including financial prediction, government intelligence, healthcare analytics, recommendation systems, or business review analysis [3]. This process begins with data collection and preprocessing, followed by feature extraction and selection, and finally classification using predictive models [4]. Numerous studies have contributed improvements at each stage of this pipeline across different languages. Most research has focused on the English language, which remains the most widely studied language in sentiment analysis. However, Arabic sentiment analysis has received increasing attention, despite several linguistic challenges, including the diversity of dialects derived from modern standard arabic (MSA), rich morphological structures, and greater linguistic complexity compared with English. These characteristics introduce additional challenges in capturing semantic consistency and increase the difficulty of modeling sentiment compared to morphologically simpler languages such as English. Moreover, most existing studies tend to focus on a single language, often overlooking multilingual contexts where multiple languages coexist. For example, in Morocco, Arabic is the primary language, while French and English are widely used as foreign languages in digital communication and online reviews. Therefore, evaluating sentiment analysis models across linguistically diverse datasets is essential to assess their generalization capabilities.

In the context of representing textual data in a numerical form, early approaches relied on statistical representations such as bag-of-words (BoW) and term frequency–inverse document frequency (TF–IDF), which capture word frequency but fail to represent semantic relationships between words. Subsequent advances introduced word embeddings, including Word2Vec, GloVe, and FastText, which provide distributed representations capturing semantic similarity. More recently, contextual embedding models based on pre-trained transformers, such as bidirectional encoder representations from transformers (BERT) and generative pre-trained transformer (GPT), have significantly improved the representation of textual data by incorporating contextual information. Nevertheless, effectively capturing the semantic meaning of entire sentences remains a challenging task. Once textual representations are obtained, they must be integrated with appropriate classification models to achieve accurate sentiment prediction. These models typically belong to several families, including machine learning algorithms such as support vector machine (SVM), random forest (RF), and decision tree (DT); deep learning architectures such as recurrent neural network (RNN), long-short-term memory (LSTM), and gated recurrent units (GRU); and more recently, models based on transformer architectures trained through fine-tuning. Another critical challenge that significantly affects sentiment classification performance is class imbalance, where the distribution of sentiment labels is skewed toward a majority class [5]. This imbalance can bias models toward the dominant class, reducing their ability to correctly identify minority sentiments. To address this issue, various data balancing and augmentation techniques have been proposed, including methods such as SMOTE and ADASYN, which generate synthetic samples to increase the representation of minority classes. However, these techniques typically rely on interpolation between existing samples and do not always guarantee the generation of realistic or semantically coherent new examples. In high-dimensional

contextual embedding spaces, such interpolation-based methods may fail to capture complex non-linear data distributions, potentially leading to suboptimal or misleading synthetic samples. Despite the growing body of work on sentiment analysis, existing studies have not sufficiently addressed the interaction between high-dimensional contextual embeddings and advanced data balancing techniques within a unified framework.

To address these challenges, we propose a novel approach that is evaluated on both English and Arabic datasets. This work is driven by the need to better understand how generative data augmentation and ensemble learning strategies can be effectively integrated with contextual sentence embeddings to improve sentiment classification performance. The proposed approach leverages Sentence-BERT to generate contextual sentence-level embeddings that capture the semantic meaning of the entire sentence. To mitigate the impact of class imbalance, we employ a Generative Adversarial Network to generate realistic synthetic samples within the embedding space, thereby improving the representation of minority classes. Finally, a soft voting ensemble classifier is used to combine the predictions of multiple machine learning models. In particular, this study aims to investigate whether WGAN-based data generation produces more representative samples than traditional resampling techniques, and whether soft voting provides advantages over alternative ensemble strategies within a contextual embedding framework. The main objectives of this study are summarized as follows:

- We employ five imbalanced datasets, including two English datasets and three Arabic datasets, to evaluate the proposed approach in a multilingual sentiment analysis setting.
- We generate sentence-level embeddings using SBERT, which capture the semantic meaning of complete sentences rather than relying on token-level representations.
- We address the class imbalance problem using a GAN-based synthetic data generation approach, and we compare its effectiveness with widely used resampling techniques such as SMOTE and ADASYN.
- We train the resulting representations using a soft voting ensemble classifier that combines several machine learning models. The performance of this ensemble is compared with individual classifiers as well as stacking-based ensemble models.

The remainder of this paper is organized as follows. Section 2 reviews related work in sentiment analysis. Section 3 describes the proposed methodology. Section 4 presents the experimental setup, analyzes the results, and discusses the findings. Finally, Section 5 concludes the paper and outlines potential directions for future research.

## 2. Related Works

Sentiment analysis studies vary across methods, processing techniques, analytical levels, and languages, with each presenting distinct characteristics and challenges. A large body of previous research has focused on the English language, often relying on traditional feature extraction techniques such as BoW [6] and TF–IDF [7], combined with classical machine learning algorithms such as RF [8], DT, k-Nearest Neighbors (KNN) or SVM. In the study conducted by [9], five machine learning classifiers are evaluated on two English-language datasets using TF–IDF representations. Experimental results demonstrate that SVM outperforms the other classifiers, particularly after applying advanced preprocessing techniques to the datasets. In the same context, ref. [10] conducted a comparative analysis of multiple feature extraction techniques. Their findings indicate that traditional machine learning models, such as SVM, Naïve Bayes (NB), and Maximum Entropy (ME), generally outperform lexicon-based approaches, particularly when TF–IDF and n-gram features are employed. Another comparative study presented by [11] demonstrates the superior performance of TF–IDF-based representations compared to lexicon-based approaches when

combined with classical machine learning classifiers such as the Support Vector Classifier (SVC), achieving a best classification accuracy of 71%. The Arabic language has also been the focus of extensive research in sentiment analysis. Ref. [12] analyzed the sentiment of Arabic tweets related to COVID-19 using BoW representations and reported that a Multinomial NB classifier achieved the best performance compared to other state-of-the-art approaches. In another study [13], the combination of TF–IDF features with Logistic Regression (LR) demonstrated strong performance among classical machine learning models, achieving an accuracy of 82%. However, contextual embedding models based on AraBERT outperformed all approaches, including machine learning and deep learning models, reaching an accuracy of 88%. Although TF-IDF and BoW are widely used, they fail to capture semantic relationships between words, particularly in short texts where contextual information is limited, leading to sparse and less informative representations. These limitations of traditional text representation techniques in capturing semantic relationships between words have motivated a shift toward more expressive embedding-based approaches. As a result, numerous studies in both Arabic and English sentiment analysis have increasingly adopted word embedding techniques, such as Word2Vec [14], GloVe [15], and FastText [16] or contextual embeddings such as BERT [17] or Efficiently Learning an Encoder that Classifies Token Replacements Accurately (ELECTRA) [18], often comparing their performance with statistical representations. In this context, ref. [19] proposed a comprehensive multilingual comparative framework for sentiment analysis, evaluating TF–IDF, Word2Vec, FastText, and BERT embeddings across balanced English, French, and Arabic datasets. Their experimental results confirmed the effectiveness of TF–IDF when combined with ensemble classifiers (EC) in English datasets, outperforming other feature representations. In contrast, contextual embeddings based on BERT demonstrated superior performance on Arabic datasets, surpassing all other feature extraction techniques, while other word embedding methods also achieved strong results across all evaluated languages. Another study [20] proposed a BERT-based framework combined with deep learning classifiers, including convolutional neural network (CNN), RNN, and bidirectional long short-term memory (BiLSTM), trained on a merged corpus of six English-language datasets. This approach was motivated by the need to overcome the limitations of traditional representations such as BoW and Word2Vec, and to effectively capture bidirectional contextual information using BERT. Experimental results demonstrated that BERT-based models consistently outperform Word2Vec-based representations, achieving classification accuracies of up to 93%. However, another study [21] adopted a framework based on GRU combined with FastText word embeddings and evaluated it on an Arabic dataset. The experimental results demonstrated strong performance, achieving an accuracy exceeding 83%. The authors of [22] proposed an approach based on Arabic-ELECTRA embeddings combined with a stacked BiLSTM and Bidirectional GRU (BiGRU) architecture enhanced by an attention mechanism. The model was evaluated on three Arabic sentiment analysis datasets and achieved state-of-the-art performance, with a best accuracy of 96.77%. Focusing on the use of contextual embeddings to enhance semantic text representations, ref. [23] employed the Robustly optimized BERT approach (RoBERTa) for feature extraction and a LSTM network as a deep learning classifier to model long-range dependencies. The proposed framework was evaluated on three English-language datasets and demonstrated superior performance compared to classical machine learning and recurrent neural baselines. In particular, the model achieved its best result on the IMDB dataset, reaching an accuracy of 92.96%. In addition, recent studies such as [24] have demonstrated that fine-tuned transformer models, such as CAMeLBERT, can achieve strong performance on Arabic sentiment analysis tasks, particularly when trained end-to-end on large annotated datasets. However, such approaches require substantial computational resources and large labeled datasets

for effective training. Although contextual embeddings improve semantic representation, their effectiveness depends on how well they are integrated with downstream tasks such as classification and data balancing.

Training models across multiple datasets, regardless of whether they are balanced or imbalanced, presents the challenge of ensuring fair class representation during learning. To address this issue, it becomes necessary to balance the class distributions so that each class contributes equally to the learning process. Accordingly, numerous studies explored techniques that either reduce the size of majority classes or augment minority classes to mitigate class imbalance. Ref. [25] conducted an in-depth study on the impact of class imbalance on model performance in sentiment analysis. Their work investigated two resampling strategies, namely Random Under-Sampling (RUS) [26] and SMOTE [27]. The experimental results from the Arabic and English datasets demonstrated that oversampling-based methods, when combined with contextual representations derived from BERT and stacking ensemble models, lead to significant improvements in sentiment classification performance. Another study [28], although not directly focused on sentiment analysis, addresses the issue of class imbalance by applying SMOTE in combination with a stacking ensemble incorporating XGBoost for default risk prediction. The findings highlight the importance of balanced class distributions in improving overall model quality and predictive performance. However, the near-perfect accuracy reported in this study raises concerns about potential overfitting or data leakage. To compare the effectiveness of resampling techniques, this study [29] conducted a comparative analysis of SMOTE and ADASYN [30] in combination with TF–IDF features and evaluated their performance using multiple classification algorithms. The results indicated that SMOTE outperformed ADASYN in this experimental setting. However, SMOTE and ADASYN rely on linear interpolation between minority samples, which may lead to the generation of unrealistic synthetic samples that do not accurately reflect the underlying data distribution, particularly in high-dimensional embedding spaces where data distributions are often non-linear. These limitations motivate the use of generative models such as WGAN, which aim to learn the underlying data distribution rather than relying on simple interpolation.

To summarize this section, Table 1 provides an overview of the reviewed studies, highlighting the different feature extraction techniques employed in the literature. Most of these approaches rely on token-level representations without fully capturing the semantic meaning of entire sentences. The table also outlines the data balancing and augmentation strategies adopted in each study. In contrast to traditional methods that duplicate existing samples, our approach generates synthetic reviews using a generative model. Furthermore, the table specifies the languages addressed in prior work, whereas our study focuses on both Arabic and English datasets.

**Table 1.** Overview of related works.

| Ref | Language | Features | Balancing | Classifier | Accuracy (%) |
|---|---|---|---|---|---|
| [9] | English | TF-IDF, BoW | - | SVM | 83.67 |
| [11] | English | BoW, TF-IDF, Word2Vec | - | SVC | 71 |
| [12] | Arabic | BoW, TF-IDF | SMOTE | Multinomial NB | 91 |
| [13] | Arabic | TF-IDF, BERT | - | AraBERT | 88 |
| [19] | Multilingual | Mixed embeddings | SMOTE | EC, SVM | 98.6 (EN), 94.1 (AR) |
| [20] | English | BoW, Word2Vec, BERT | - | BERT | 93 |
| [21] | Arabic | Word2Vec, FastText | - | IAN-BiGRU | 83.98 |
| [22] | Arabic | AraELECTRA | SMOTE | Stacked BiLSTM-GRU | 96.77 |
| [23] | English | RoBERTa | - | RoBERTa-LSTM | 92.96 |
| [25] | Multilingual | BERT | SMOTE | Stacking Model | 94.0 (EN), 94.2 (AR) |
| [28] | P2P dataset | LightGBM | SMOTE | Stacking XGBoost | 99.98 |
| [29] | English | TF-IDF, Word2Vec | SMOTE, ADASYN | Extra Trees | 98.26 |

## 3. Materials and Methods

In this study, we propose an approach based on SBERT embeddings, which generate fixed-length sentence representations. SBERT encodes each sentence into a dense semantic vector that captures its overall meaning. To address class imbalance, we applied a GAN-based synthetic data generation strategy to augment the minority class within the training set. A soft voting ensemble learning method is employed, combining three base classifiers by averaging their prediction probabilities to produce the final sentiment decision. The general structure of the proposed approach is illustrated in Figure 1.

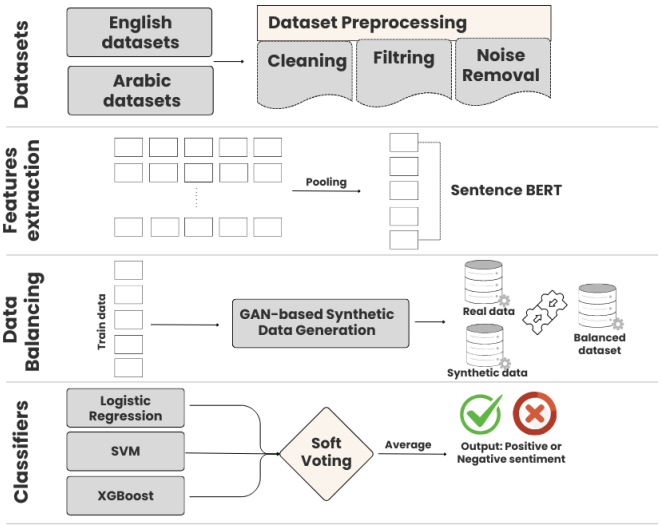



**Figure 1.** Structure of the proposed approach.

### *3.1. Data Processing*

The datasets used in this study go through a multi-stage preprocessing pipeline. The process begins with the removal of the neutral sentiment class, which represents the minority across all datasets, in order to focus on a binary sentiment classification task by retaining only positive and negative instances. This choice aims to better analyze class imbalance. In business intelligence applications, identifying negative feedback is often more critical than distinguishing neutral opinions. For Arabic datasets, specific language-dependent preprocessing steps are applied. The diacritics, called Tashkil, which correspond to vowel markings, are removed, as well as the elongation characters, called Tatwil. In addition, repeated characters used for emphasis are reduced to a single occurrence, and variant letter forms are normalized to their base representations. For both Arabic and English datasets, URLs, HTML tags, and stopwords are removed. However, negation terms are deliberately preserved to maintain the semantic orientation and polarity of sentences. These terms are identified using a predefined list of common Arabic negation particles (e.g., ‘la’, ‘lam’, ‘ma’, ‘laysa’) and are retained whether they appear as standalone tokens or attached prefixes. Stemming and lemmatization are not applied in this work. This decision is motivated by the use of Sentence SBERT, a contextual embedding model that encodes sentences at a global semantic level rather than relying on individual lexical forms. Applying stemming or lemmatization could distort the contextual meaning of sentences and degrade the quality of the resulting embeddings, unlike traditional representations such as BoW or TF–IDF.

### *3.2. Features Extraction*

Feature extraction is a crucial step in the sentiment analysis pipeline, as it involves transforming raw textual data into numerical representations that can be processed by

machine learning and deep learning models. Several approaches have been proposed to generate such representations, ranging from traditional statistical techniques, such as TF-IDF and BoW, to more advanced word embedding methods, including GloVe, FastText, and Word2Vec. However, with the advent of transformer-based models, particularly BERT and its variants, embedding generation has reached a new level of effectiveness by capturing rich contextual and semantic relationships between words. Despite their strong contextual modeling capabilities, BERT-based models are not ideally suited for sentence-level sentiment analysis, as they produce token-level representations rather than a single fixed-size vector representing the entire sentence. This limitation is especially evident in multilingual sentiment analysis tasks using models such as AraBERT for Arabic language and BERT for English language. To address this issue, Sentence-BERT is employed, as a technique to generate semantically sentence-level embeddings.

### Sentence-BERT

Sentence-BERT (SBERT), introduced by [31], is an extension of the BERT architecture specifically designed to generate fixed-length vector representations for entire sentences. Unlike the original BERT model, which produces contextualized embeddings at the token level, SBERT encodes a complete sentence into a single, dense, and semantically meaningful vector that captures the overall meaning of the text. This makes SBERT suitable for our task, as it reduces the need for extensive linguistic preprocessing while preserving rich semantic information. SBERT adopts a Siamese network architecture, allowing it to learn sentence embeddings by optimizing similarity measures between sentence pairs, typically using cosine similarity. SBERT builds upon a transformer backbone such as BERT or RoBERTa and introduces an additional pooling layer applied to the transformer outputs, the pooling operation aggregates the token-level representations into a single 1 × dimension sentence vector as illustrated in Figure 2. Several pooling strategies can be applied, including cls pooling, max pooling, and mean pooling, with mean pooling being the most commonly used approach. In the context of sentiment analysis, SBERT transforms each sentence into a numerical vector that effectively captures its semantic and emotional content. Consequently, sentences expressing similar sentiments are mapped to nearby vectors in the embedding space, whereas sentences with opposing sentiments are positioned farther apart.

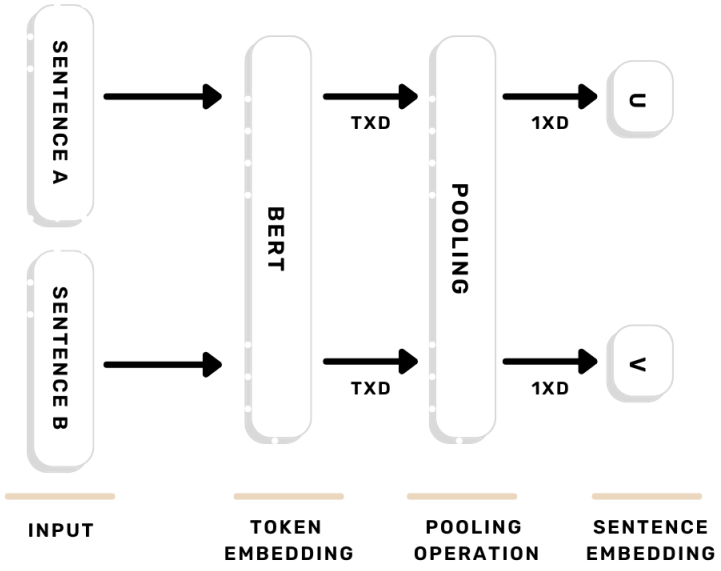



**Figure 2.** SBERT embeddings architecture.

### 3.3. Data Balancing

Addressing the issue of class imbalance is an important step in improving the performance of sentiment classification models. The datasets employed in this study exhibit a significant imbalance among sentiment classes, which can lead learning algorithms to focus on the majority class while neglecting minority classes. As a result, models often achieve high overall accuracy but suffer from low recall for minority classes and exhibit overfitting toward the dominant class. Several techniques have been proposed to traits the class imbalance problem. Under-sampling methods reduce the number of majority class instances; however, this approach may result in the loss of valuable information. Conversely, oversampling techniques increase the representation of minority classes by duplicating existing samples, which can introduce redundancy and increase the risk of overfitting. More advanced methods, such as SMOTE and ADASYN, generate synthetic samples by interpolating between minority class instances. In this work, these techniques are compared with a generative adversarial network (GAN)-based synthetic data generation approach, which is adopted in this study.

#### GAN-Based Synthetic Data Generation

Generative Adversarial Network (GAN) [32] is a class of generative models widely used for synthetic data generation. A GAN consists of two neural networks trained simultaneously in an adversarial framework: a generator G and a discriminator D. The discriminator is trained to distinguish between real samples drawn from the true data distribution and synthetic samples produced by the generator, while the generator is trained to produce synthetic data that closely resemble real samples from the target domain. This adversarial process can be viewed as a competitive game in which the discriminator aims to correctly classify real and fake samples, whereas the generator seeks to deceive the discriminator by generating increasingly realistic synthetic data. During training, each network iteratively updates its behaviours in response to the performance of its opponent. In this work, we adopt a variant of GAN known as the Wasserstein Generative Adversarial Network (WGAN) [33], which addresses several limitations of the classical GAN framework. Unlike standard GAN that output probabilities, WGAN provides a real-valued critic score that approximates the Wasserstein distance between real and generated data distributions, resulting in more stable training and improved convergence behaviour. The WGAN is applied to the embeddings of the minority class extracted from the training data using SBERT, and the generated samples are added to the training data before classifier training. Specifically as shown in Figure 3, the generator takes minority-class SBERT embeddings as input and maps them through two fully connected layers followed by an output layer with a tanh activation function, to produce synthetic embeddings with the same dimensionality as the real ones. The discriminator, also referred to as the critic, receives either real or synthetic embeddings and outputs a scalar score using two fully connected layers. The critic is optimized to maximize the difference between the scores assigned to real and synthetic embeddings, while enforcing a gradient penalty to satisfy the Lipschitz continuity constraint. Training is performed by updating the critic three times for each generator update, using the Adam optimizer with a learning (lr $= 10^{-4}$). The model is trained for 10 epochs on the minority-class embeddings. After convergence, the trained generator is used to produce synthetic minority embeddings until class balance between minority and majority classes is achieved. Figures 4 and 5 illustrate a two-dimensional projection of SBERT embeddings before and after WGAN-based augmentation for the Hotel Reviews and HTL datasets. The comparison between the distributions before and after balancing shows that the generated synthetic samples densify the minority-class region while preserving the overall geometric structure of the embedding space. Critic score

analysis suggests that synthetic embeddings approximate the distribution of real ones, although with slightly lower scores. However, this does not provide a direct evaluation of semantic realism. Therefore, in this study, their quality is primarily validated through improved downstream classification performance.

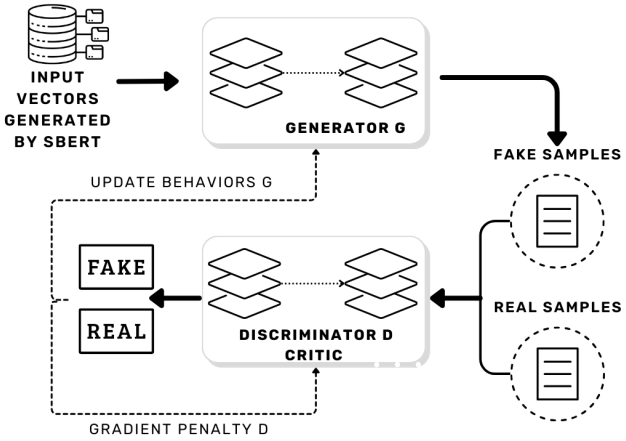
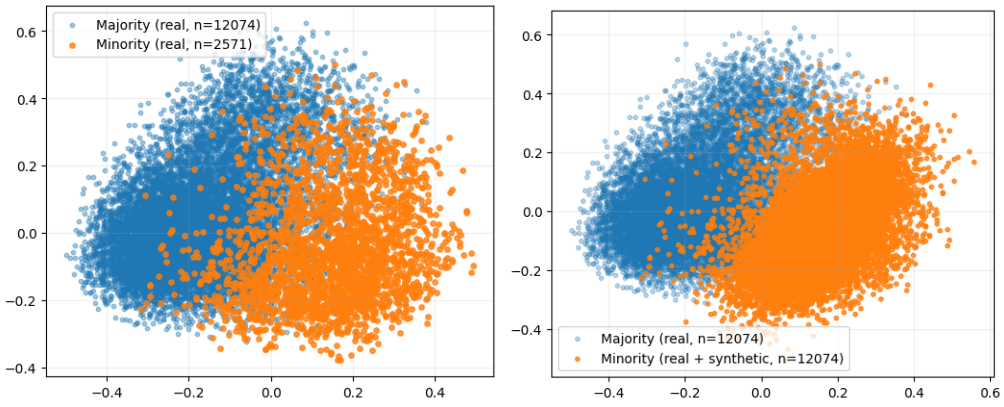
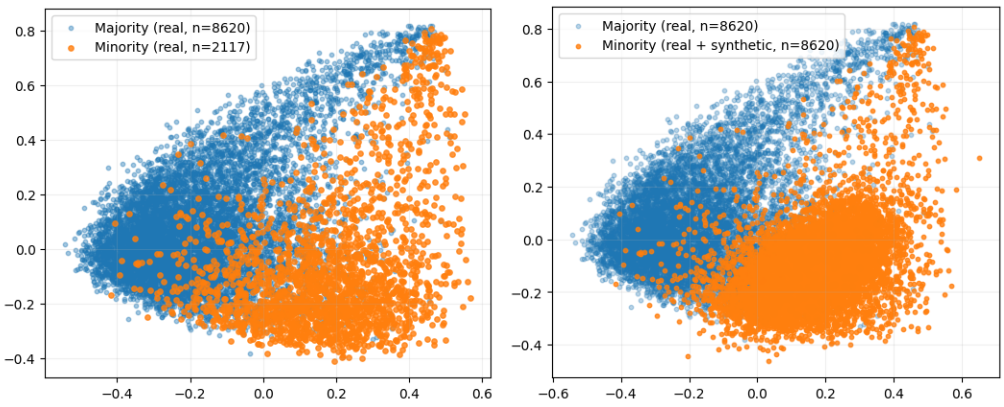



**Figure 3.** Wasserstein GAN model for synthetic minority embeddings.




(**a**) (**b**)

**Figure 4.** SBERT embeddings before (**a**) and after (**b**) WGAN-based balancing for the TripAdvisor Hotel Reviews dataset.




(**a**) (**b**)

**Figure 5.** SBERT embeddings before (**a**) and after (**b**) WGAN-based balancing for the HTL dataset.

## 4. Experiments Analysis and Results

This section presents the datasets used in this study, the classification models employed, and the evaluation metrics adopted to assess model performance. Then it reports the experimental results obtained on both English and Arabic datasets, followed by a discussion of these findings.

### *4.1. Datasets*

The proposed approach is evaluated on English and Arabic datasets related to the business intelligence domain, which were collected from online platforms featuring hotel and restaurant reviews. This domain consistency is intentional, as domain-specific vocabulary in business intelligence contexts enables more coherent and task-relevant representations. Table 2 provides an overview of the characteristics of these datasets.

**Table 2.** Overview of the datasets.

| Dataset Name | Language | Dataset Size | |
|---|---|---|---|
| | | Positive | Negative |
| TripAdvisor Hotel | English | 15,093 | 3214 |
| Yelp Res Reviews | English | 15,330 | 2497 |
| HTL | Arabic | 10,775 | 2647 |
| RES | Arabic | 8030 | 2675 |
| SemEval-2016 | Arabic | 7705 | 4556 |

#### 4.1.1. English Datasets

The English datasets used in this work are collected from online hotel review platforms, and two benchmark datasets are used for the evaluation of the experiment.

- TripAdvisor Hotel Reviews [34]: This is an English-language benchmark dataset collected from the TripAdvisor platform and contains more than 20,000 customer reviews covering various aspects of hotel quality. Each review was originally associated with a rating ranging from 1 to 5. To address a binary sentiment classification, the ratings are mapped into two sentiment categories: reviews with ratings of 4 and 5 are labeled as positive, while those with ratings of 1 and 2 are labeled as negative. Reviews with a neutral rating of 3 are excluded from the analysis. After this transformation, the resulting dataset becomes imbalanced, which allows the application and evaluation of the data balancing technique proposed in this work.
- Yelp Restaurant Reviews [35]: This dataset comprises more than 19,800 reviews collected from over 45 restaurants on the Yelp platform. Each review was annotated with a rating ranging from 1 to 5. Following the same procedure applied to the TripAdvisor dataset, ratings of 4 and 5 are mapped to positive sentiment, while ratings of 1 and 2 are mapped to negative sentiment. Reviews with a rating of 3 are ignored to maintain a binary classification. The resulting dataset contains 17,827 reviews with an imbalanced class distribution.

#### 4.1.2. Arabic Datasets

In this study, we evaluate the proposed approach using three Arabic datasets collected from online hotel and restaurant review platforms.

- HTL dataset [36]: The Arabic hotel dataset was collected from the TripAdvisor platform and contains 15,572 reviews labeled as positive, negative, and neutral. The neutral category constitutes the minority proportion of the data. For the purposes of this study, we restrict the analysis to positive and negative sentiments to address the binary classification, as neutral reviews are less informative for sentiment analysis in

business intelligence applications. The resulting dataset remains imbalanced, which enables the application and evaluation of different balancing techniques.

- RES dataset [36]: The Arabic restaurant dataset consists of Arabic customer reviews collected from more than 4500 restaurants listed across two review platforms: the Qaym startup and the TripAdvisor website. The dataset contains 10,970 reviews, with positive and negative sentiments constituting the majority of the samples. Following the same preprocessing strategy adopted for the first dataset, only positive and negative sentiment categories are retained.
- SemEval-2016 [37]: The SemEval-ABSA16 dataset comprises consumer reviews of more than 15,000 hotels collected from platforms such as Booking.com and TripAdvisor. It contains 13,113 reviews annotated with positive, negative, and neutral sentiment labels, with the neutral class representing the minority. In line with the preprocessing strategy adopted for the previous datasets, the neutral category is excluded. Although this dataset was originally designed for aspect-level sentiment analysis, it is also used in this study for sentence-level sentiment classification.

### *4.2. Baselines Classifiers*

In this subsection, we introduce the classifiers used in this study. We adopt an ensemble learning strategy for the classification stage, specifically using a soft voting method. The ensemble is composed of three machine learning classifiers: LR, SVM, and XGBoost. Deep learning models such as RNN, LSTM, and GRU are not considered in this work, as these architectures are designed to process sequential and token-level representations, which require token-by-token embeddings. In contrast, our approach operates on global sentence-level representations generated by SBERT, which produces fixed-size dense vectors that capture the overall semantics of each sentence. Furthermore, we do not rely on end-to-end fine-tuning of transformer-based architectures, such as BERT or the Arabic specific AraBERT, for model training. Instead, SBERT is used exclusively as a feature extractor to obtain sentence-level semantic embeddings. These embeddings are then augmented using a GAN-based approach in a continuous latent space before being fed into the ensemble classifier. Accordingly, the adoption of a soft voting ensemble learning approach is driven by the sentence-level representation of the data.

#### 4.2.1. Ensemble Classifier

Ensemble classifiers aim to combine multiple base models to produce a final prediction that is generally more accurate than that of a single model. By aggregating the predictions of different classifiers, ensemble methods reduce individual model errors and improve generalization performance. Several ensemble learning strategies exist. First, bagging [38] trains multiple instances of the same model on different subsets of the training data. Boosting [39] builds models sequentially, where each subsequent model focuses on correcting the errors made by the previous ones. Stacking [40] combines the predictions of multiple base learners by training a meta-classifier on their outputs. In this study, we adopt a voting-based ensemble classifier [41], which integrates the predictions of heterogeneous base classifiers. The final decision is obtained by aggregating the individual model outputs through a voting mechanism.

#### 4.2.2. Soft Voting Technique

One of the most widely used ensemble classification strategies is voting-based ensemble learning, which combines the predictions of independently trained base models to generate a final decision. Voting ensembles can be implemented using either hard voting or soft voting mechanisms. Hard voting, also known as majority voting, assigns the final class label based on the class that receives the highest number of votes from the

individual classifiers. In contrast, soft voting, also called weighted voting, aggregates the class probability estimates produced by each classifier. The final prediction is obtained by averaging these probabilities and selecting the class with the highest aggregated score. In this study, we adopt the soft voting ensemble approach, which integrates the predictions of three base classifiers. The selected classifiers were chosen for their complementary properties, combining linear and non-linear decision boundaries as well as probabilistic outputs. The ensemble model is designed to reduce model variance and leverage complementary decision boundaries from different classifiers.

- Logistic Regression (LR): is a supervised machine learning model commonly used for binary classification tasks. It estimates the probability that an input instance belongs to a given class by computing a linear combination of the input features and then applying a non-linear sigmoid function. This transformation maps the output to a value in the range $[0, 1]$, which can be interpreted as a probability of membership in the class [42].
- Support Vector Machine (SVM): is a supervised learning algorithm for classification tasks that aims to identify the optimal separating hyperplane between classes by maximizing the margin between them. This margin corresponds to the maximum distance between the hyperplane and the closest data points from each class [43].
- eXtreme Gradient Boosting (XGBoost): is an ensemble learning algorithm based on the gradient boosting framework. It builds a strong predictive model by training a sequence of weak learners, DT in our study, in a sequential manner. Each newly added tree is optimized to correct the errors made by the previous models by minimizing a loss function through gradient-based optimization [44].

### *4.3. Parameter Settings*

The proposed approach is implemented through multiple sequential stages, each associated with specific parameter settings. The process begins with feature extraction using SBERT. In this study, we employ the multilingual SBERT model to transform each sentence into a fixed-length vector of 768 dimensions. In particular, the paraphrase-multilingual-mpnet-base-v2 model was selected due to its strong performance in multilingual semantic similarity tasks and its ability to generate high-quality sentence embeddings across different languages. The generated embeddings are subsequently used as input to a generative data augmentation stage based on a WGAN with Gradient Penalty. As high-dimensional spaces make it more difficult to learn stable data distributions. To mitigate this issue, the WGAN generator projects the 768-dimensional embeddings into a 512-dimensional latent space before reconstructing them back to the original dimension using fully connected layers. The WGAN is trained using the Adam optimizer with a learning rate of 0.0001 for 10 epochs and a batch size of 256. The number of training epochs was set to 10 based on prior studies. This choice is supported by additional experiments conducted with 15 and 25 epochs, which did not yield noticeable improvements in performance and convergence. Furthermore, limiting training duration helps reduce the risk of overfitting and mode collapse while promoting stable convergence. For model evaluation, the datasets used in this work were divided using a hold-out validation strategy, with 80% of the data allocated for training and 20% reserved for testing. Several classifiers are employed in this work, each configured with specific hyperparameters. LR is trained using a maximum of 3000 iterations with the Limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) solver. The SVM classifier is implemented with a linear kernel and includes feature standardization. The XGBoost classifier is configured with 400 DT and a maximum tree depth of 6. All experiments are conducted on the Google Colab platform using Python version 3.12.12. The complete experimental configuration and parameter settings are summarized in Table 3.

**Table 3.** Experimental parameter settings.

| Parameter | Value |
|---|---|
| SBERT | |
| Sentence embedding model | paraphrase-multilingual-mpnet-base-v2 |
| Embedding dimension | 768 |
| Pooling strategy | Mean pooling |
| WGAN | |
| Training data | Minority class only |
| Input scaling | Min–Max scaling to $[-1, 1]$ |
| Generator architecture | Dense (512)–Dense (512) |
| Critic architecture | Dense (512)–Dense (256) |
| Optimizer | Adam |
| Learning rate | $1 \times 10^{-4}$ |
| Adam parameters | $\beta_1 = 0.0$, $\beta_2 = 0.9$ |
| Gradient penalty coefficient ($\lambda$) | 10 |
| Batch size | 256 |
| Number of epochs | 10 |
| Critic updates step | 3 |
| Classification Models | |
| Logistic Regression | L2 regularization, lbfgs solver, $C = 1.0$, max_iter = 3000 |
| SVM | Linear kernel, $C = 1.0$ |
| XGBoost | Number of trees = 400, maximum depth = 6, learning rate = 0.05 |

### *4.4. Evaluation Metrics*

The performance of the proposed approach and the baseline models is evaluated using multiple standard classification metrics, including accuracy, precision, recall, F1-score, the area under curve value, the ROC curve and the Matthews correlation coefficient. These evaluation metrics are widely used in classification tasks and provide a comprehensive assessment of model performance.

Accuracy, precision, recall, and F1-score are standard evaluation metrics derived from the confusion matrix and are used to assess classification performance. Accuracy provides an overall measure of the proportion of correctly predicted instances across all classes. Precision focuses on the positive predictions by quantifying the proportion of correctly predicted positive samples among all samples predicted as positive. Recall measures the model's ability to identify positive instances by computing the proportion of correctly predicted positive samples among all actual positive samples. The F1-score represents the harmonic mean of precision and recall. The mathematical formulations of these metrics are presented as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

The area under the curve (AUC) is an important evaluation metric that reflects how effectively a model distinguishes between the two classes: positive and negative sentiments in this study. The AUC score ranges between 0 and 1, with higher values indicating better discriminative performance.

The receiver operating characteristic (ROC) curve provides a visual representation of the trade-off between two key rates: the true positive rate (TPR), which corresponds to recall, and the false positive rate (FPR), which measures the proportion of negative instances incorrectly classified as positive. The ROC curve is particularly important for evaluating the performance of classification models in scenarios where class imbalance is present, as is the case in this study. The mathematical formulations of the TPR and FPR are presented as follows:

$$\text{TPR} = \frac{TP}{TP + FN} \tag{5}$$

$$\text{FPR} = \frac{FP}{FP + TN} \tag{6}$$

Another important evaluation metric is the Matthews correlation coefficient (MCC), which summarizes the quality of binary classification models. MCC quantifies the correlation between the true and predicted class labels and is especially well suited for imbalanced classification problems. The mathematical formulation of MCC is defined as follows:

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{7}$$

where $TP$, $TN$, $FP$, and $FN$ denote true positives, true negatives, false positives, and false negatives, respectively.

### *4.5. Experimental Findings*

In this section, we present the experimental results of the proposed approach described in Section 3. The model is evaluated on five datasets, including two English datasets and three Arabic datasets. Its performance is compared with several baseline models, especially LR, SVM, XGBoost, MLP, and a stacking ensemble method. Comparisons with individual classifiers are included to demonstrate ensemble effectiveness, while comparisons between ensemble methods, offer a more meaningful evaluation. In addition, other comparative methods are also discussed in this subsection, in order to assess the effectiveness of the proposed approach. The performance results of the proposed approach on the English datasets are reported in Table 4, which presents three evaluation metrics: Accuracy, F1-score, and MCC. The proposed approach achieves the best performance on the TripAdvisor Hotel Reviews dataset across all three metrics, obtaining 95.08% accuracy, 97.02% F1-score, and 82.96% MCC. These results outperform all baseline models, particularly the stacking ensemble and the MLP classifier, which achieve accuracies of 94.18% and 94.19%, and F1-scores of 96.53% and 96.13%, respectively. In contrast, LR and XGB yield the lowest performance on this dataset, both achieving an accuracy of 93.73%. A similar trend is observed on the Yelp Restaurant Reviews dataset. The proposed model achieves the highest accuracy of 95.74%, surpassing the SVM model with 94.46% accuracy and the MLP classifier with 94.49%. In contrast, the stacking ensemble achieves the lowest accuracy at 93.63%, although the difference remains relatively small. In terms of F1-score and MCC, the proposed approach also achieves the best results, with an F1-score of 97.53% and an MCC of 82.09%, outperforming other models such as MLP, which obtains an F1-score of 96.40% and an MCC of 80.37%. In summary, all evaluated models were trained using the same feature extraction method and the same data balancing strategy. Therefore, the observed improvement can be attributed primarily to the effectiveness of the proposed soft voting ensemble model combined with SBERT embeddings and WGAN-based data augmentation. The soft voting ensemble classifier enhances the predictive capability of the framework by aggregating the strengths of multiple machine learning models including LR, SVM, and XGB. By leveraging the complementary decision patterns of these classifiers. This

combination explains the superior performance of the proposed approach compared to individual classifiers as well as more complex architectures such as stacking ensembles and MLP models.

The performance of the proposed architecture on the Arabic datasets is presented in Table 5. The results show that the proposed model consistently outperforms the baseline classifiers across all datasets. Specifically, it achieves accuracies of 94.41%, 86.92%, and 91.89% on the HTL, RES, and SemEval-2016 datasets, respectively. These results surpass those obtained by the competing models, including XGB, which achieves 93.12% on the HTL dataset, MLP, which reaches 86.22% on the RES dataset, and SVM, which records 90.99% on the SemEval-2016 dataset. In contrast, the lowest performance is observed for LR on the HTL and SemEval-2016 datasets, with accuracies of 92.71% and 90.48%, respectively, while the stacking ensemble yields the lowest accuracy on the RES dataset with 82.81%. A similar trend can be observed for the F1-score, which reflects the balance between precision and recall. The proposed model achieves the highest F1-scores across all Arabic datasets, reaching 96.52% on HTL, 91.74% on RES, and 93.53% on SemEval-2016, again outperforming all baseline models. Furthermore, the MCC, which is particularly informative for imbalanced datasets, confirms the superiority of the proposed approach. The model achieves MCC values of 82.39%, 63.76%, and 82.66% on the HTL, RES, and SemEval-2016 datasets, respectively, representing the highest scores among all evaluated methods. Similar to the English datasets, the superior performance is attributed to the effective combination of SBERT embeddings, WGAN-based synthetic minority data generation, and soft voting ensemble classification. SBERT provides rich sentence-level semantic representations, while WGAN generates realistic synthetic embeddings for the minority class, improving class balance without duplicating existing samples. Finally, the soft voting ensemble aggregates the probabilistic predictions of heterogeneous classifiers. The overall performance gains are interpreted as the combined effect of SBERT, WGAN, and soft voting model. The synergy of these components leads to consistent improvements across all evaluation metrics, including accuracy, precision, recall, F1-score, and MCC, on all five datasets. The observed improvements remain meaningful for imbalanced sentiment classification, particularly for enhancing minority class detection in business intelligence applications. Although the performance gains are modest, they are consistently observed across multiple validation settings, including different train–test splits and k-fold cross-validation.

**Table 4.** Performance comparison across English datasets.

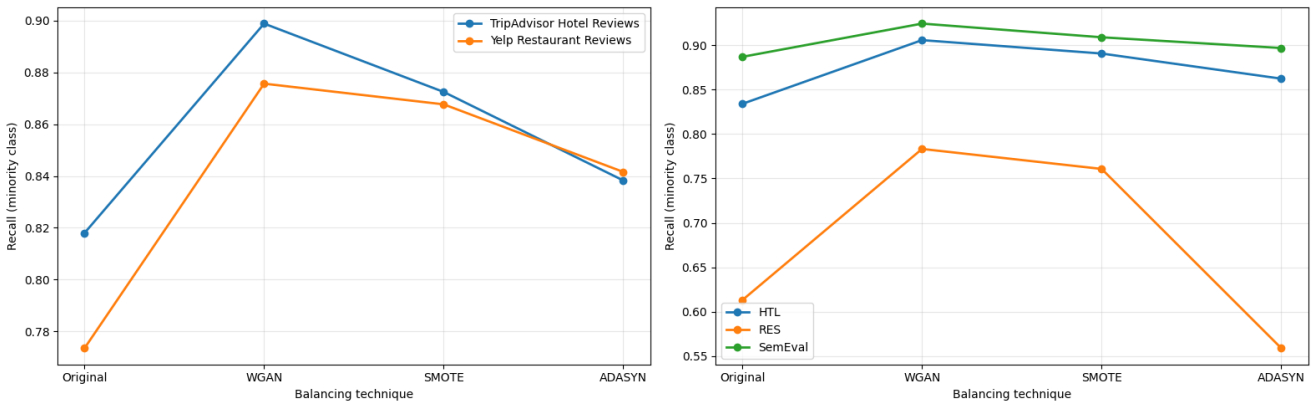
| Classifier | TripAdvisor Hotel Reviews | | | Yelp Restaurant Reviews | | |
|---|---|---|---|---|---|---|
| | **Acc** | **F1** | **MCC** | **Acc** | **F1** | **MCC** |
| LR | 93.73 | 95.80 | 81.85 | 94.34 | 96.30 | 80.44 |
| SVM | 94.03 | 96.01 | 82.39 | 94.46 | 96.38 | 80.45 |
| XGB | 93.73 | 95.81 | 81.68 | 94.23 | 96.24 | 79.23 |
| MLP | 94.19 | 96.13 | 82.75 | 94.49 | 96.41 | 80.37 |
| Stacking | 94.18 | 96.53 | 78.97 | 93.63 | 96.40 | 71.07 |
| Proposed approach | 95.08 | 97.02 | 82.96 | 95.74 | 97.53 | 82.09 |

To evaluate the effectiveness of the data balancing strategy used in the proposed approach, we compare the WGAN-based synthetic data generation method with SMOTE and ADASYN. All methods employ the same SBERT sentence embeddings and the same soft voting ensemble classifier. Furthermore, consistent with the WGAN, SMOTE, and ADASYN are applied to the training set after the data split, ensuring that the test set remains completely unseen. Synthetic samples are generated until class balance is achieved by matching the number of minority samples to that of the majority class. The results presented in Table 6 show that the WGAN consistently improves performance across all

five datasets. On the TripAdvisor Hotel Reviews and Yelp Restaurant Reviews datasets, the proposed WGAN achieves relative accuracy improvements of 0.79% and 0.98%, respectively, compared with SMOTE, and 1.55% and 1.57% compared with ADASYN. Similar improvements are achieved on the Arabic datasets. On the HTL, RES, and SemEval-2016 datasets, the WGAN-based method improves performance by 1.48%, 2.19%, and 1.45%, respectively, compared with SMOTE, and by approximately 2% compared with ADASYN. Among the evaluated techniques, ADASYN consistently produces the lowest performance across all datasets. These results highlight the advantage of generating more realistic synthetic minority representations using WGAN. Unlike SMOTE, which creates synthetic samples through linear interpolation between neighboring instances, and ADASYN, which extends SMOTE by focusing on more difficult regions of the feature space, WGAN learns the underlying distribution of the minority class in the continuous embedding space. This capability becomes particularly important when working with high-dimensional sentence embeddings produced by SBERT. By modeling the distribution of minority embeddings directly, WGAN generates more coherent and semantically meaningful synthetic samples, leading to improve classification performance. As illustrated in Figure 6, this approach leads to a significant improvement in minority-class recall and overall classification performance across all evaluated datasets. For the English datasets, the WGAN-based approach consistently outperforms SMOTE, ADASYN, and the original imbalanced datasets without any balancing technique. The same applies to the Arabic datasets, where WGAN increases minority-class recall. In contrast, the original unbalanced datasets exhibit the lowest recall values, followed by ADASYN and then SMOTE.

**Table 5.** Performance comparison across Arabic datasets.

| Classifier | HTL | | | RES | | | SemEval-2016 | | |
|---|---|---|---|---|---|---|---|---|---|
| | Acc | F1 | MCC | Acc | F1 | MCC | Acc | F1 | MCC |
| LR | 92.71 | 95.07 | 80.36 | 85.08 | 89.90 | 61.58 | 90.48 | 92.23 | 81.73 |
| SVM | 93.12 | 95.36 | 81.12 | 84.31 | 89.19 | 54.34 | 90.99 | 92.78 | 80.81 |
| XGB | 93.41 | 95.51 | 81.55 | 85.94 | 90.79 | 61.28 | 90.50 | 92.40 | 79.74 |
| MLP | 93.12 | 95.30 | 81.93 | 86.22 | 91.06 | 61.51 | 90.72 | 92.35 | 81.43 |
| Stacking | 93.08 | 95.35 | 80.85 | 82.81 | 88.95 | 51.16 | 90.75 | 92.77 | 80.03 |
| Proposed approach | 94.41 | 96.52 | 82.39 | 86.92 | 91.74 | 63.76 | 91.89 | 93.53 | 82.66 |




(**a**) (**b**)

**Figure 6.** Minority class recall across balancing techniques for (**a**) English datasets and (**b**) Arabic datasets.

**Table 6.** Performance comparison of SMOTE, ADASYN, and WGAN using a fixed soft voting ensemble (Accuracy, %).

| Classifier | English Datasets | | Arabic Datasets | | |
|---|---|---|---|---|---|
| | TripAdvisor Hotel Reviews | Yelp Restaurant Reviews | HTL | RES | SemEval-2016 |
| SMOTE with Voting | 94.29 | 94.76 | 92.93 | 84.73 | 90.44 |
| ADASYN with Voting | 93.53 | 94.17 | 92.48 | 83.42 | 89.93 |
| Proposed Approach | 95.08 | 95.74 | 94.41 | 86.92 | 91.89 |

To further evaluate the discriminative capability of the models, Figures 7 and 8 present the ROC curves and AUC scores for each baseline model across all evaluated datasets. The results indicate that the proposed soft voting ensemble consistently outperforms the individual classifiers as well as the stacking ensemble method. This is related to the WGAN-based balancing strategy, which reduces the bias toward the majority class and improves the TPR without excessively increasing the FPR. In addition, the soft voting stabilizes the final predictions by aggregating the probabilistic outputs of multiple classifiers. In contrast, the SVM classifier yields the lowest ROC–AUC values on the English datasets, while the stacking ensemble exhibits the weakest performance on the Arabic datasets.

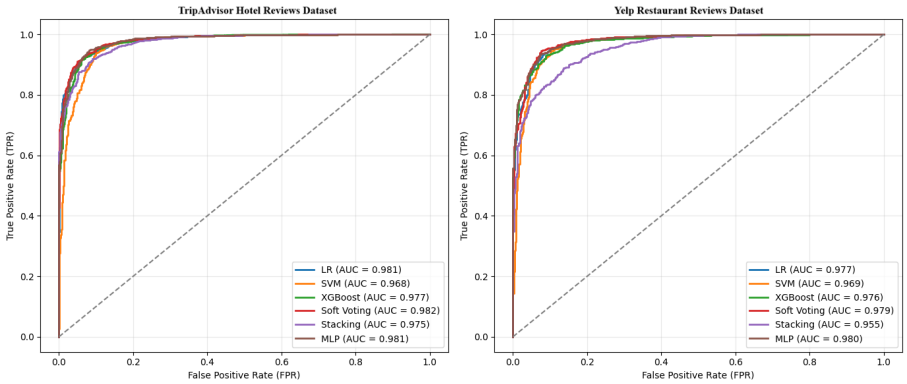
In addition, we evaluated the proposed model using different validation strategies as shown in Table 7, in order to assess its stability and generalization capability. The initial experiments were conducted using an 80:20 train–test split, which was also used for all baseline models in the previous results. To verify the robustness of the approach, we compared this configuration with other partitioning strategies, including 90:10 and 70:30 fixed train-test splits, as well as 5-fold and 10-fold cross-validation. The results show that the 80:20 split generally provides the best performance across most datasets, particularly for the TripAdvisor Hotel Reviews, RES, and SemEval-2016 datasets. In contrast, the 90:10 split achieves slightly higher performance on the Yelp Restaurant Reviews and HTL datasets. Performance variations across validation strategies, particularly on the Yelp dataset, can be attributed to differences in class distribution and data partitioning. Overall, the performance differences across the various validation strategies remain relatively small, indicating that the proposed approach is stable and does not depend on a specific data partition.

**Table 7.** Performance comparison across English and Arabic datasets using fixed train–test splits and K-fold validation (Accuracy, %).

| Dataset | Fixed Train–Test | | | K-Fold | |
|---|---|---|---|---|---|
| | 80:20 | 90:10 | 70:30 | K = 5 | K = 10 |
| TripAdvisor Hotel Review | 95.08 | 94.81 | 95.07 | 95.04 | 95.07 |
| Yelp Restaurant Reviews | 95.74 | 97.42 | 97.06 | 95.25 | 95.38 |
| HTL | 94.41 | 95.23 | 94.31 | 94.73 | 94.84 |
| RES | 86.92 | 85.62 | 85.08 | 85.55 | 85.66 |
| SemEval-2016 | 91.89 | 90.95 | 91.87 | 91.26 | 91.25 |

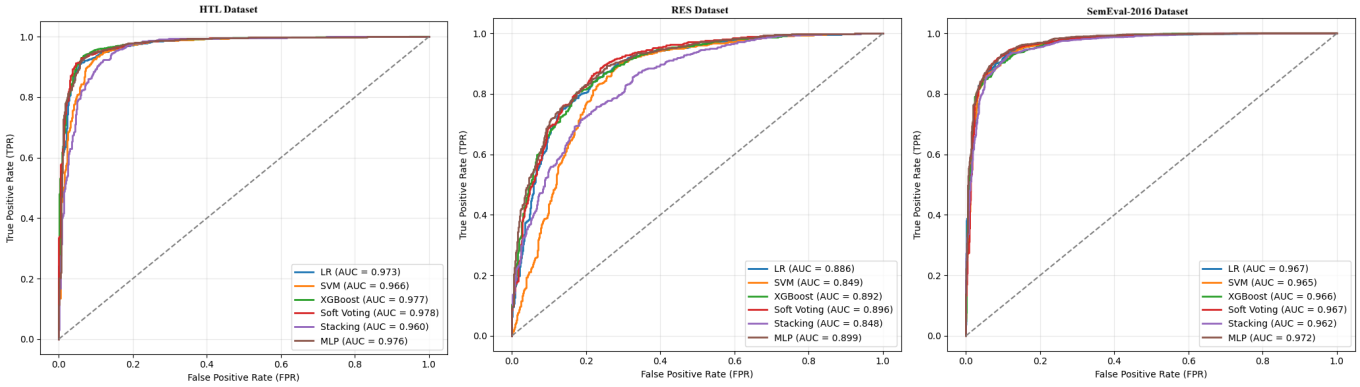
Figures 9 and 10 present a comparison of the classifiers used in this study in terms of training and inference time across all datasets. Inference time refers to the time required to generate predictions on new data. In the English datasets, the soft voting ensemble requires approximately 12 times more training time than XGBoost, the fastest individual classifier, while stacking is more than 2.5 times slower than soft voting. Similarly, in the Arabic datasets, the soft voting approach is approximately 8 times slower than XGBoost, and stacking remains more than 2.5 times slower than soft voting. The results show that

ensemble-based models incur the highest computational cost, both during training and inference time, due to their multi-level architecture involving multiple base learners and in the case of stacking, a meta-learner. The soft voting ensemble represents the second most computationally expensive approach. In contrast, individual classifiers such as LR and SVM exhibit significantly lower training and inference times. These findings highlight the trade-off between model complexity and computational efficiency, where the proposed approach achieve improved predictive performance at the expense of increased computational cost.



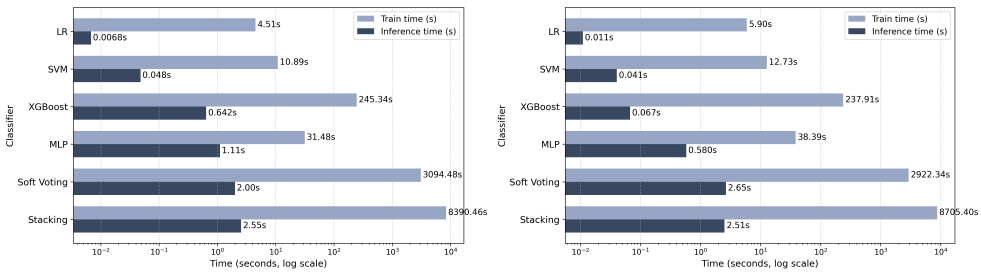

(**a**) (**b**)

**Figure 7.** ROC–AUC curves obtained using the proposed approach on two English sentiment analysis datasets: (**a**) TripAdvisor Hotel reviews. (**b**) Yelp Restaurant reviews.



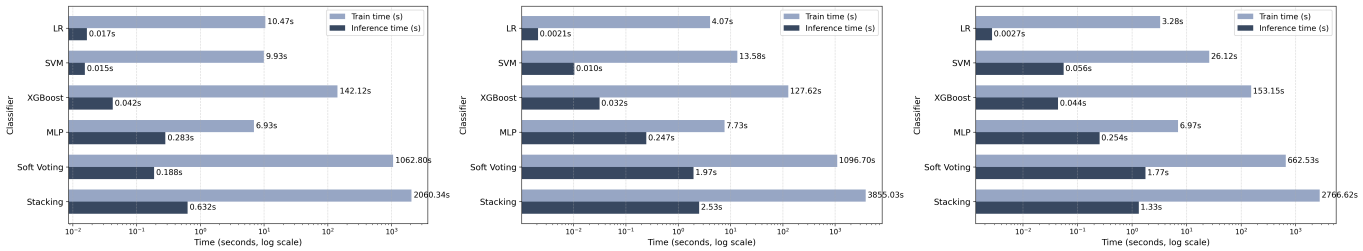

(**a**) (**b**) (**c**)

**Figure 8.** ROC–AUC curves obtained using the proposed approach on three Arabic sentiment analysis datasets: (**a**) HTL dataset. (**b**) RES dataset. (**c**) SemEval-2016.




(**a**) (**b**)

**Figure 9.** Training and inference time comparison on TripAdvisor Hotel Reviews dataset (**a**) and Yelp Restaurant Reviews dataset (**b**) .




(**a**) (**b**) (**c**)

**Figure 10.** Training and inference time comparison on HTL dataset (**a**), RES dataset (**b**) and SemEval-2016 dataset (**c**).

## 5. Conclusions

The growing demand for effective sentiment analysis models has significantly accelerated research in this field. Early studies primarily focused on the English language, but recent efforts have expanded to include more linguistically complex languages such as Arabic. In parallel, research has increasingly addressed each stage of the sentiment analysis pipeline, including preprocessing, feature extraction, and classification in order to develop more accurate models.

In this study, we focus on two languages, English and Arabic, within the context of business intelligence. To evaluate the proposed approach, we employ five datasets, including two English datasets and three Arabic datasets. The datasets go through a several preprocessing steps before applying SBERT, a contextual embedding technique that generates fixed-length sentence representations capable of capturing the overall semantic meaning of each review. To address the issue of class imbalance, we apply a WGAN-based synthetic data generation to generate realistic synthetic samples for the minority class in the embedding space. The resulting balanced representations are then used to train a soft voting ensemble classifier that combines several machine learning algorithms, including LR, SVM, and XGBoost. Experimental results demonstrate the effectiveness of the proposed approach across multiple evaluation metrics, including accuracy, precision, recall, F1-score, MCC, and ROC–AUC score. The performance of the proposed method is further validated through comparisons with baseline models, alternative balancing techniques such as SMOTE and ADASYN, and different validation strategies. These include fixed train–test splits (80:20, 90:10, and 70:30) as well as cross-validation with 5 and 10 folds, confirming the stability of the proposed model. Our proposed approach outperforms all baselines models acroos both English and Arabic datasets, demonstrating that the integration of SBERT-based contextual sentence embeddings, WGAN-driven synthetic minority data generation and soft voting ensemble learning effectively captures sentence level semantic information while addressing class imbalance, thereby improving sentiment classification performance.

Despite these advances, research in sentiment analysis continues to evolve. Future work will further explore improvements in feature extraction, data balancing, and data augmentation techniques. While LLMs such as GPT achieve strong performance, they require high computational resources and limited interpretability; the proposed approach remains a more efficient and controllable alternative, with future work exploring hybrid integration. Future work will also extend the proposed approach to multi-class sentiment classification, including the neutral class, to improve its applicability to real-world scenarios.

In addition, we will extend this study by fine-tuning additional transformer models and evaluating them using both the proposed approach and alternative methods to provide a more comprehensive analysis. Moreover, we plan to investigate heuristic optimization

techniques for feature selection and incorporate explainable artificial intelligence (XAI) methods, including SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME), to enhance the interpretability and transparency of sentiment analysis models.

**Author Contributions:** Conceptualization, H.J. and S.E.H.; methodology, H.J.; software, H.J., S.T. and M.-A.E.H.; validation, H.J., S.E.H. and M.-A.E.H.; formal analysis, H.J. and S.A.; investigation, H.J., S.T. and S.A.; resources, H.J. and S.E.H.; data curation, H.J.; writing—original draft preparation, H.J.; writing—review and editing, H.J.; visualization, H.J., S.T. and S.E.H.; supervision, S.E.H. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Data Availability Statement:** The English datasets used in this study are publicly available via Kaggle at https://www.kaggle.com/datasets/andrewmvd/trip-advisor-hotel-reviews, accessed on 2 January 2026 and at https://www.kaggle.com/datasets/farukalam/yelp-restaurant-reviews, accessed on 20 February 2026. The Arabic datasets used in this paper are publicly available in Github at https://github.com/hadyelsahar/large-arabic-sentiment-analysis-resouces, accessed on 18 December 2025 and via Huggingface at https://huggingface.co/datasets/srinivasbilla/semeval-2016-absa-reviews-arabic, accessed on 19 December 2025.

**Acknowledgments:** We used writing support tools such as Grammarly and LanguageTool to help identify and correct grammatical and lexical issues in some parts of the manuscript. In addition, Google Translate was occasionally used for the translation of certain words or expressions when needed. However, we would like to clearly state that no generative AI tools were used for the scientific generation of the article, including the methodology, experiments, analysis, results, or conclusions, all of which were entirely developed and validated by the authors.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Tan, K.L.; Lee, C.P.; Lim, K.M. A Survey of Sentiment Analysis: Approaches, Datasets, and Future Research. *Appl. Sci.* **2023**, *13*, 4550. [CrossRef]
2. Zhang, L.; Liu, B. Sentiment Analysis and Opinion Mining. In *Synthesis Lectures on Human Language Technologies*; Springer: Berlin/Heidelberg, Germany, 2012. [CrossRef]
3. Mao, Y.; Liu, Q.; Zhang, Y. Sentiment analysis methods, applications, and challenges: A systematic literature review. *J. King Saud Univ. Comput. Inf. Sci.* **2024**, *36*, 102048. [CrossRef]
4. Birjali, M.; Kasri, M.; Hssane, A.B. A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowl. Based Syst.* **2021**, *226*, 107134. [CrossRef]
5. Altalhan, M.; Algarni, A.; Alouane, M.T.H. Imbalanced Data Problem in Machine Learning: A Review. *IEEE Access* **2025**, *13*, 13686–13699. [CrossRef]
6. Zhang, Y.; Jin, R.; Zhou, Z.H. Understanding bag-of-words model: A statistical framework. *Int. J. Mach. Learn. Cybern.* **2010**, *1*, 43–52. [CrossRef]
7. Salton, G.; Buckley, C. Term-Weighting Approaches in Automatic Text Retrieval. *Inf. Process. Manag.* **1988**, *24*, 513–523. [CrossRef]
8. Bahrawi, N. Sentiment Analysis Using Random Forest Algorithm-Online Social Media Based. *J. Inf. Technol. Its Util.* **2019**, *2*, 29–33. [CrossRef]
9. Sunitha, P.; Joseph, S.; Akhil, P.V. A Study on the Performance of Supervised Algorithms for Classification in Sentiment Analysis. In *Proceedings of the TENCON 2019—2019 IEEE Region 10 Conference (TENCON)*; IEEE: New York, NY, USA, 2019; pp. 1351–1356. [CrossRef]
10. Devika, M.; Sunitha, C.; Ganesh, A. Sentiment Analysis: A Comparative Study on Different Approaches. *Procedia Comput. Sci.* **2016**, *87*, 44–49. [CrossRef]
11. Qi, Y.; Shabrina, Z. Sentiment analysis using Twitter data: A comparative application of lexicon- and machine-learning-based approach. *Soc. Netw. Anal. Min.* **2023**, *13*, 31. [CrossRef]
12. Albahli, S. Twitter Sentiment Analysis: An Arabic Text Mining Approach Based on COVID-19. *Front. Public Health* **2022**, *10*, 966779. [CrossRef]

13. Abdellah, A.E.; Cherrat, E.M.; Ouahi, H.; Bekkar, A. Sentiment Analysis from Texts Written in Standard Arabic and Moroccan Dialect Based on Deep Learning Approaches. *Int. J. Comput. Digit. Syst.* **2024**, *16*, 447–458. [CrossRef] [PubMed]
14. Mikolov, T.; Chen, K.; Corrado, G.S.; Dean, J. Efficient Estimation of Word Representations in Vector Space. In Proceedings of the International Conference on Learning Representations, Scottsdale, AZ, USA, 2–4 May 2013.
15. Pennington, J.; Socher, R.; Manning, C.D. GloVe: Global Vectors for Word Representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2014. [CrossRef]
16. Bojanowski, P.; Grave, E.; Joulin, A.; Mikolov, T. Enriching Word Vectors with Subword Information. *Trans. Assoc. Comput. Linguist.* **2016**, *5*, 135–146. [CrossRef]
17. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the North American Chapter of the Association for Computational Linguistics, Minneapolis, MN, USA, 2–7 June 2019. [CrossRef]
18. Clark, K.; Luong, M.T.; Le, Q.V.; Manning, C.D. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In Proceedings of the International Conference on Learning Representations, Addis Ababa, Ethiopia, 30 April 2020. [CrossRef]
19. Jakha, H.; Houssaini, S.E.; Houssaini, M.A.E.; Ajjaj, S.; Hadir, A. Optimizing Sentiment Analysis in Multilingual Balanced Datasets: A New Comparative Approach to Enhancing Feature Extraction Performance with ML and DL Classifiers. *Appl. Syst. Innov.* **2025**, *8*, 104. [CrossRef]
20. Bello, A.; Ng, S.C.; Leung, M.F. A BERT Framework to Sentiment Analysis of Tweets. *Sensors* **2023**, *23*, 506. [CrossRef]
21. Abdelgwad, M.M.; Soliman, T.H.A.; Taloba, A.I.; Farghaly, M.F. Arabic aspect based sentiment analysis using bidirectional GRU based models. *J. King Saud Univ. Comput. Inf. Sci.* **2021**, *34*, 6652–6662. [CrossRef]
22. Jakha, H.; Houssaini, S.E.; Houssaini, M.A.E.; Ajjaj, S.; Kafi, J.E. S2BA-AraELECTRA: A stacked BiLSTM-BiGRU with attention mechanism and contextual embeddings from AraELECTRA for enhanced Arabic sentiment classification in business intelligence. *Int. J. Inf. Technol.* **2025**, 1–16. [CrossRef]
23. Tan, K.L.; Lee, C.P.; Anbananthen, K.S.M.; Lim, K.M. RoBERTa-LSTM: A Hybrid Model for Sentiment Analysis With Transformer and Recurrent Neural Network. *IEEE Access* **2022**, *10*, 21517–21525. [CrossRef]
24. Aljomah, F.; Aldhafeeri, L.; Alfadel, M.; Alshahrani, S.; Abbas, Q.; Alhumoud, S. Enhancing Arabic Sentiment Analysis with Pre-Trained CAMeLBERT: A Case Study on Noisy Texts. *Comput. Mater. Contin.* **2025**, *84*, 5317–5335. [CrossRef]
25. Habbat, N.; Hicham, N.; Anoun, H.; Hassouni, L. Sentiment analysis of imbalanced datasets using BERT and ensemble stacking for deep learning. *Eng. Appl. Artif. Intell.* **2023**, *126*, 106999. [CrossRef]
26. Kubát, M.; Matwin, S. Addressing the Curse of Imbalanced Training Sets: One-Sided Selection. In Proceedings of the International Conference on Machine Learning, Nashville, Tennessee, 8–12 July 1997.
27. Chawla, N.; Bowyer, K.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
28. Muslim, M.A.; Nikmah, T.L.; Pertiwi, D.A.A.; Subhan; Jumanto; Dasril, Y.; Iswanto. New model combination meta-learner to improve accuracy prediction P2P lending with stacking ensemble learning. *Intell. Syst. Appl.* **2023**, *18*, 200204. [CrossRef]
29. Umer, M.; Sadiq, S.; Missen, M.M.S.; Hameed, Z.; Aslam, Z.; Siddique, M.A.; Nappi, M. Scientific papers citation analysis using textual features and SMOTE resampling techniques. *Pattern Recognit. Lett.* **2021**, *150*, 250–257. [CrossRef]
30. He, H.; Bai, Y.; Garcia, E.A.; Li, S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*; IEEE: New York, NY, USA, 2008; pp. 1322–1328. [CrossRef]
31. Reimers, N.; Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv* **2019**, arXiv:1908.10084. [CrossRef]
32. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; Volume 27. [CrossRef]
33. Arjovsky, M.; Chintala, S.; Bottou, L. Wasserstein Generative Adversarial Networks. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017. [CrossRef]
34. Alam, M.H.; Ryu, W.J.; Lee, S. Joint multi-grain topic sentiment: Modeling semantic aspects for online reviews. *Inf. Sci.* **2016**, *339*, 206–223. [CrossRef]
35. Public Domain. Yelp Restaurant Reviews Dataset. 2017. Available online: https://www.kaggle.com/datasets/farukalam/yelp-restaurant-reviews (accessed on 20 February 2026).
36. ElSahar, H.; El-Beltagy, S.R. Building Large Arabic Multi-domain Resources for Sentiment Analysis. In Proceedings of the Conference on Intelligent Text Processing and Computational Linguistics, Cairo, Egypt, 14–20 April 2015. [CrossRef]

37. Al-Smadi, M.; Qawasmeh, O.; Talafha, B.; Al-Ayyoub, M.; Jararweh, Y.; Benkhelifa, E. An enhanced framework for aspect-based sentiment analysis of Hotels' reviews: Arabic reviews case study. In *Proceedings of the 2016 11th International Conference for Internet Technology and Secured Transactions (ICITST)*; IEEE: New York, NY, USA, 2016; pp. 98–103. [CrossRef]
38. Breiman, L. Bagging Predictors. *Mach. Learn.* **1996**, *24*, 123–140. [CrossRef]
39. Freund, Y.; Schapire, R.E. A decision-theoretic generalization of on-line learning and an application to boosting. In Proceedings of the European Conference on Computational Learning Theory, Jerusalem, Israel, 17–19 March 1997.
40. Wolpert, D.H. Stacked generalization. *Neural Netw.* **1992**, *5*, 241–259. [CrossRef]
41. Kittler, J.; Hatef, M.; Duin, R.P.W.; Matas, J. On Combining Classifiers. *IEEE Trans. Pattern Anal. Mach. Intell.* **1998**, *20*, 226–239. [CrossRef]
42. Peng, C.Y.J.; Lee, K.L.; Ingersoll, G.M. An Introduction to Logistic Regression Analysis and Reporting. *J. Educ. Res.* **2002**, *96*, 3–14. [CrossRef]
43. Evgeniou, T.; Pontil, M. Support Vector Machines: Theory and Applications. In Proceedings of the Machine Learning and Its Applications, Williamstown, MA, USA, 28 June–1 July 2001. [CrossRef]
44. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; Association for Computing Machinery: New York, NY, USA, 2016. [CrossRef]