Improving Customer Review Analysis through Hybrid Evolutionary SVM Method using Imbalanced DataSet

Alekhya Rayala[1,] Ramesh Ponnala[2]

[1]MCA Student, Chaitanya Bharathi Institute of Technology (A), Gandipet, Hyderabad, Telangana State, India

[2]Assistant Professor, Department of MCA, Chaitanya Bharathi Institute of Technolog(A), Gandipet, Hyderabad, Telangana State, India

ABSTRACT: The quantity of customer evaluations for restaurants and the influence of online media on restaurant operations are both growing.Customers and those who make decisions in this industry rely heavily on these reviews as their primary information source. As a result, as customer feedback is regarded as the ultimate assessment of any restaurant's general quality. It might have an impact on the performance of the restaurant industry. The sentiments underlying these reviews can be analysed and predicted using Sentiment Analysis (SA). This work proposes a hybrid approach that combines the Support Vector Machine algorithm with Particle Swarm Optimisation and other oversampling techniques to handle the problem of imbalanced data. SVM is employed as a machine learning classification technique to predict the sentiments of user reviews by optimising the dataset, which consists of diverse reviews. In order to produce an optimised dataset and solve the dataset's imbalance issue, four different oversampling techniques, namely SMOTE, SVM-SMOTE, ADASYN and borderline-SMOTE, were investigated. This study demonstrates that, for various versions of the datasets, the proposed PSO-SVM approach performs other classification techniques

Keywords – Sentiment analysis, SVM, PSO, SMOTE, oversampling, feature extraction, features weighting

I. INTRODUCTION

over the past few decades, more people are engaging in online activities such social media communications-commerce, blogging, and surfing. According to a recent trend, customers now prefer to read reviews of a product before purchasing it.[7] As in today's overly socially connected society, individuals place more trust in authentic customer reviews than in flashy advertising advertisements. As it becomes simpler to choose a decent restaurant for a certain cuisine, this trend has been very beneficial for the restaurant's patrons and client support. As a result, it requires restaurant owners to gather and keep records of consumer reviews on social media platforms.

By incorporating customer suggestions, sentiment analysis of customer reviews also aids in improving the overall customer experience.[1] The amount of product reviews available has significantly expanded as a result of the widespread use of social networks and applications, and the demand for automated ways to gather and analyze these evaluations has also increased. These techniques are necessary to expedite and enhance the decision-making process.

By analyzing implicit attitudes and the hidden sentiments in comments, SA can be used to predict user's opinions about a variety of issues.[9] Analyzing people's sentiments, opinions, appraisals, attitudes, evaluations and emotions towards such entities as businesses, products, services, individuals, topics, issues, events and their attributes, as presented online via text, video and other means of communication.

This study suggests an evolutionary method for examining people's reactions to reviews of restaurants written in Arabic. Additionally, this work used a hybrid evolutionary strategy, combining the PSO algorithm with several oversampling methods in order to automatically identify the In the previous few years, social media websites' popularity has grown dramatically. Due to the widespread usage of the internet sentiment in the customers' remarks, along with the SVM algorithm. To address the issue of imbalance in the dataset, four alternative oversampling strategies are used.[3] By determining the optimal feature weights and k value for the oversampling technique, the applied evolutionary algorithm also contributes to reducing the time and effort required to modify the parameters and optimize the classification, leading to superior performance metrics. After applying the SVM method to categorize the weighted oversampled data, the outcomes will be evaluated using G-mean. The individual variables are then optimized using the Particle Swarm Optimizer algorithm to produce a higher G-mean.

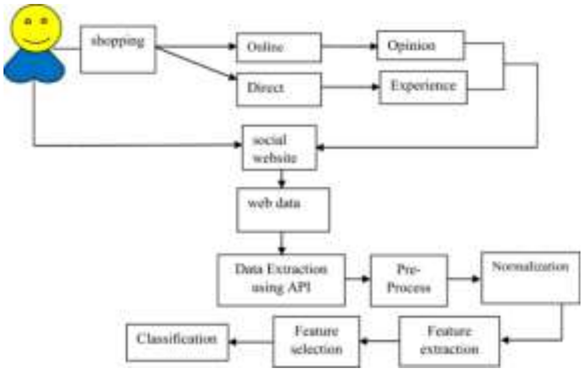

Figure.1: The system's overall functionality

II.LITERATURE REVIEW

Naimul Hossain[1]in their study addressed that businesses are gradually leaning towards online delivery services, and customer reviews are now used to evaluate restaurants' overall quality

Sindhu Hegde[2]discussed the problems that arise while starting a new restaurant business.. To maximize the profit, they first determine the restaurant characteristics or aspects that patrons are most drawn to, and then offer such amenities and services Finally, since location has a significant impact on a restaurant's ability to succeed, they believe that knowing the area around a location is essential.

Leen Muteb Alharbi[3]in their study discuss that in today's world, social media is crucial. People can share their opinions and views regarding the goods that are offered on e-commerce websites, which are frequently referred to as an assessment or judgement.. The best outcomes were attained using Support Vector Machine, Logistic Regression, and Random Forest.
Minh-Hao Nguyen[4] in their study discusss the issue of aspect-based sentiment analysis that has drawn more attention from scholars. The objective is to gather insightful data on the topics stated in user comments. The three subtasks of word extraction, aspect detection, and polarity detection can be applied to this issue.

Oman Somantri[5]discussed about the Consumer reviews or opinions on restaurants that serve culinary food will result in information that may be used to help people make decisions about where they'll get these kinds of foods. The best classification method was used to create a text mining-based sentiment analysis model utilising the review text data.

Kanwal Zahoor[6]in their study addressed the use of social networking sites that has significantly expanded during the past several years. Social media platforms are used by people to express their opinions on nearly any topic. Customer input is crucial for organizations, and since social media is such a strong platform, it can be leveraged to develop and improve company chances.

Marwan Al Omari[7] in their study suggests a logistic regression method along with term and inverse document frequency (TF*IDF) for categorizing Arabic sentiment in reviews of services in the country of Lebanon. Public services including hotels, restaurants, stores, and others are the subject of reviews. They manually gathered reviews from Zomato and Google, totaling 3916 reviews.

Anu Taneja[8] discussed that as a result of advancements in the web, research on user behaviour is becoming more and more popular. Check-ins on Facebook are among the finest ways to engage with users' places of interest out of several research areas. Such research is certainly advantageous for services like location recommendations. The main goal of this study project is to comprehend, examine, and recommend restaurants and locations based on user.

Maria Habib[9] discussed that emailing systems need to be secured from spam because it is one of the main forms of Internet communication and poses a serious hazard to both individual users and businesses.Because of this problem, it is imperative to create more precise and efficient spam detection models for emailing platforms.

Anjana Gosain[10] in their study addressed that classifier's goal is to divide items in a data set into one or more groups according to their features. In practical applications, classifiers are used on sets of data that are out of whack The performance of conventional classification algorithms is negatively impacted by unbalanced data sets.

III METHODOLOGY

A. DESCRIPTION AND COLLECTION OF DATA

The dataset used in this study is detail reviews left by customers of various Jordanian restaurants. Data has been collected from Jeeran, a popular social network for Arabic ratings.[3] This website provides a comparison and evaluation platform for the top establishments and services in the Arab world since 2010, including cafes, hotels, restaurants, and public services. Reviews of such establishments can offer important insight to those who decide on matters such as the standard of the food and service, costs, and other ambiance-related factors from the Jeeran website, almost 3000 restaurant reviews have been collected.

B.DATA PREPARATION AND LABELLING

Before being uploaded to the dataset, it is cleaned, labelled, formatted, and stemmed. By deleting symbols and special characters in dataset, the cleaning procedure is carried out. The reviewers were instructed to thoroughly examine each review and identify it according to the customer opinion. The reviewers might choose from two options for each review: negative(1) and positive(0). As a result, the review's class was determined by the choices made by the majority of reviewers.

All reviews were compiled into a CSV file, with their class label in one column and their context in the second. After labelling the dataset, formatting is initiated. First, all stop words are eliminated, such as I'm, so, that, then, very, this, and may, respectively. Stop words must be removed because they have no bearing on the text's meaning. After, through a normalization procedure, any non-Arabic letters and emojis are removed. Text normalization and stop word removal were used to remove a lot of pointless features, reducing the overall amount of extracted features and improving the feature selection procedure.

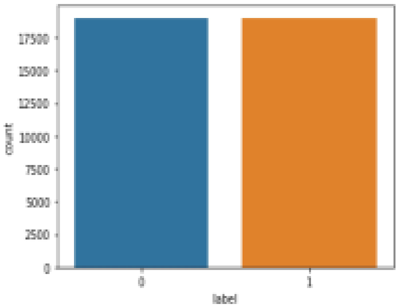

Figure2:Representation of balanced class distribution achieved through SMOTE.

C.PROPOSED SYSTEM

This study proposes a hybrid approach that combines the Support Vector Machine algorithm with Particle Swarm Optimisation and other oversampling techniques to handle the problem of imbalanced data. As a machine learning classification method, SVM is used to predict the sentiments of reviews. This is achieved by optimising the dataset, which consists of numerous reviews of various Jordanian restaurants. The information was gathered via Jeeran, a well-known social network for Arabic evaluations. Four distinct oversampling approaches were researched to create an efficient dataset and address the imbalanced issue.

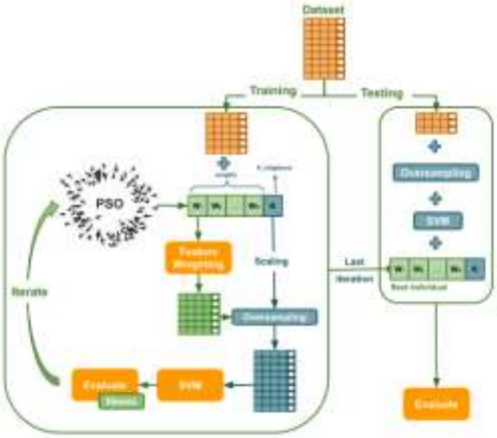

Figure.3: A visual representation showcasing the PSO-SVM approach employing oversampling techniques.[26].

IV. IMPLEMENTATION

ALGORITHMS:

SVM – SVM is a powerful supervised algorithm that works best on smaller datasets but on complex ones. SVM can be used for both regression and classification tasks, but generally, they work best in classification problems. When working with imbalanced datasets, the Hybrid Evolutionary SVM technique is intended to increase the precision of sentiment analysis in customer review analysis. It combines the strength of evolutionary methods with Support SVM to improve classification performance and optimize SVM hyperparameters.

The basic concept of the hybrid evolutionary SVM technique is to utilise evolutionary algorithms to find the best set of SVM hyperparameters [4]. Examples of these algorithms include genetic algorithms and particle swarm optimization. In comparison to conventional SVM models with default settings, the approach tries to improve classification results by fine-tuning the hyperparameters specifically for imbalanced datasets.

PSO- Particle swarm optimization (PSO) swarm intelligence algorithm was created to solve nonlinear problems in a variety of scientific and technical fields. It was inspired by how birds and fish school. PSO, a method that employs swarm intelligence to find answers. It analyses a set of potential solutions (known as a swarm), each of which is referred to as a particle, and produces a random search result.

Normal moving particles rely on two types of learning: social learning and cognitive learning. The first describes the process of learning from other particles (the outcome is saved as best), while the second describes about the process of storing the best solution that may be found.

Bi-LSTM: A layer that develops the bidirectional long-term dependencies between time steps of time series or sequence data is known as a bidirectional LSTM (BiLSTM). When you want the network to learn from the entire time series at each time step, these dependencies can be helpful. A bidirectional LSTM, often known as a biLSTM, is a sequence processing model that consists of two LSTMs, one of which receives input forward and the other of which receives it backward[6].. Additionally, their present state can be used to get their future input information.

Bi-RNN: Bidirectional recurrent neural networks (BRNN) link two concealed layers that are facing in different directions to the same output. The output layer can simultaneously receive data from previous (backwards) and future (ahead) states with this type of generative deep learning. BRNNs were developed in 1997 by Schuster and Paliwal in order to expand the network's access to input data.

For instance, because they require constant input data, multilayer perceptron (MLPs) and time delay neural network (TDNNs) have restrictions on the flexibility of their input data. Standard recurrent neural networks (RNNs) also have limitations because the information for future input cannot be accessed from the state of the network today. BRNNs, on the other hand, don't need their input data to be fixed.

Bi-GRU- is a model for processing sequences that consists of two GRUs. one processing the information forward and the other processing it backward. Only the input and forget gates are present in this neural network.

GRU: The subtype of recurrent neural network (RNN), the gated recurrent unit (GRU), occasionally outperforms long short-term memory (LSTM). GRU is quicker and requires less memory than LSTM, however LSTM is more accurate when working with datasets that contain longer sequences.

Kyunghyun Cho et al[7]. presented gated recurrent units (GRUs) as a gating technique for recurrent neural networks in 2014. Voting Classifier (LR + RF) – Voting Classifier is a machine-learning algorithm often used by Kagglers to boost the performance of their model and climb up the rank ladder.[11] Voting Classifier can also be used for real-world datasets to improve performance, but it comes with some limitations.

LSTM: Long short-term memory (LSTM) is a type of artificial neural network that is employed in deep learning and artificial intelligence. LSTM features feedback connections as opposed to typical feedforward neural networks. Such a recurrent neural network (RNN) can analyse whole data sequences, such as audio or video, in addition to single data points, like images.[9] For instance, LSTM can be used for applications like speech recognition, machine translation, robot control, unsegmented, networked handwriting recognition, video games, and healthcare.

SVM + SMOTE – Synthetic Minority Oversampling Technique (SMOTE) is a very popular oversampling method that was proposed to improve random oversampling but its behavior on high-dimensional data has not been thoroughly investigated.

OVERSAMPLINMG TECHNIQUES-The problem that frequently arises in classification challenges is when the target class label is distributed unevenly. These data can be thought of as an unbalanced dataset, which has an impact on the data mining model's training process because it will be focused mostly on the majority class, leading to bias in class predictions because the minority class's few instances may be viewed as noise or outliers. Solving data imbalance concerns is essential and should be done before classification as was done in As a result, imbalanced datasets pose major hurdles by affecting the performance of classifiers. In this regard, a variety of equilibrium strategies are used. Oversampling techniques, including SMOTE and adaptive synthetic sampling (ADASYN), can be used to group them.

V. EXPERIMENTS AND RESULTS

The results of the analysis are presented in this part, along with an assessment of the classifiers' effectiveness. The performance of the classifiers can be evaluated using the common parameters listed.

A. Accuracy

Accuracy is usually employed to evaluate a classification algorithm's performance. The number of samples which have been accurately estimated to the total anticipated samples is what is meant b y accuracy.

$Accuracy = TP + TN$Accuracy = TP + TN /TP + TN + FP + FN (1)

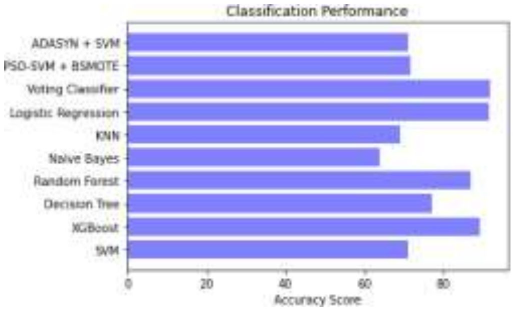

Figure4:Classification performance of accuracy score

B. Precession

Out of all the reviews which were projected to be favourable (or negative), it calculates the percentage of accurately predicted positive (or negative) reviews. The following formula is used to calculate precision, which is important for evaluating the precision of the model's positive and negative predictions.

Precession=True positives/(True Positives+False Positives) (2)

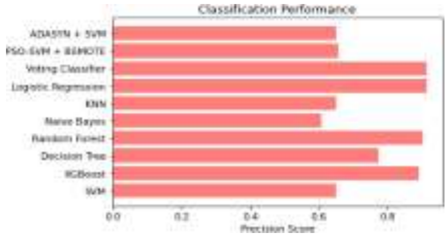

Figure5: Classification performance of Precession score

C. Recall

The recall rate is the percentage of examples that are accurately classified as positive to all examples that are classified as positive.

Recall = TP/TP + FN (3)

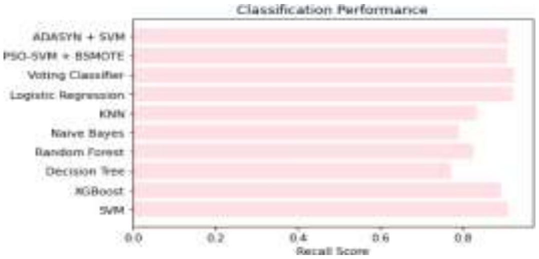

Figure6:Classification performance of Recall score

D. F1 Score

The harmonic mean of precision and recall is represented by the F-measure. It serves as a measurement tool for sentiment classification analysis.

F mean = 2*Recall*Precision/Accuracy + Recall (4)

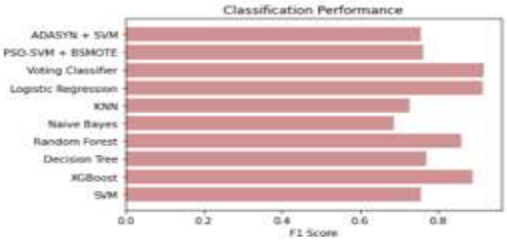

Figure 7: Classification performance of F1 Score

E. AUC

AUC (Area Under the ROC Curve): Based on the projected probabilities, it provides a measure of how well a model can distinguish between favourable and unfavourable customer evaluations.Sentiment analysis is a technique used in customer review analysis to ascertain the sentiment or opinion expressed in a particular review. The AUC score is a useful metric for evaluating how well sentiment analysis models categorize customer evaluations.

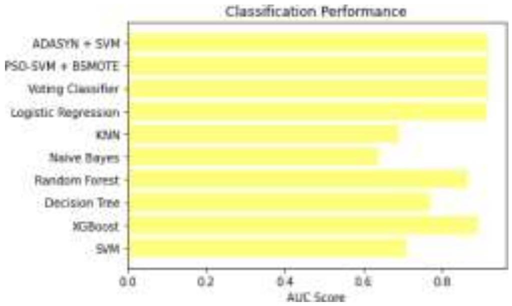

Figure8: Classification performance of AUC Score

F.G-Mean

The G-mean is a metric used to compare how well two classes performed when categorising data. In mathematics, recall-negative (RECN) and recall-positive (RECP) recollections are multiplied by the square root to get G-means.

$$G - mean = p \surd RECN \times RECP \quad (5)$$

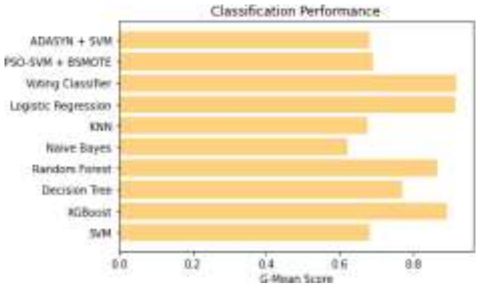

Figure 9: Classification performance of G-Mean

VI. CONCLUSION

Researchers in the field has been more interested in sentiment analysis during the past few years. Reviews of various goods and services are frequently posted online. All businesses, including restaurants, must analyse the attitudes and feedback of their customers. As a result, this study presented a novel hybrid evolutionary method that seeks to analyse consumers' perceptions of numerous eateries throughout Jordan. The information was gathered through Jeeran, a well-unbalanced data was then resolved by using oversampling techniques. In order to determine the appropriate weights and the k values of four distinct oversampling algorithms to predict the feelings of reviews, we built a hybrid optimisation technique combining PSO and SVM.

The study shows that the suggested PSO-SVM technique is efficient and performs better than the other approaches in all tested metrics (accuracy, F-measure, g-means, and AUC). More specifically, in all versions of the datasets, the PSO-SVM outperformed the regular SVM, LR, RF, DT, k-NN, and XGBoost. By applying voting classifier,91.75% accuracy was achieved.

On this data, we intend to use a variety of metaheuristic algorithms in the future. In addition, other applications can be used to forecast the tone of evaluations for different items, including those in the engineering and medical fields.

REFERENCES

[1] Y. M. Aye and S. S. Aung, ‘‘Senti-lexicon and analysis for restaurant reviews of Myanmar text,’’ Int. J. Adv. Eng., Manage. Sci., vol. 4, no. 5, Jan. 2018, Art. no. 240004.

[2] P. P. Rokade and A. K. D, ‘‘Business intelligence analytics using sentiment analysis—A survey,’’ Int. J. Electr. Comput. Eng., vol. 9, no. 1, p. 613, Feb. 2019.

[3] K. Zahoor, N. Z. Bawany, and S. Hamid, ‘‘Sentiment analysis and classification of restaurant reviews using machine learning,’’ in Proc. 21st Int. Arab Conf. Inf. Technol. (ACIT), Nov. 2020, pp. 1–6.

[4] R. Ponnala and C. R. K. Reddy, “Software Defect Prediction using Machine Learning Algorithms: Current State of the Art,” Solid State Technol., vol. 64, no. 2, 2021.

[5] M. Nakayama and Y. Wan, ‘‘The cultural impact on social commerce: A sentiment analysis on yelp ethnic restaurant reviews,’’ Inf. Manage., vol. 56, no. 2, pp. 271–279, Mar. 2019.

[6] Q. Gan, B. H. Ferns, Y. Yu, and L. Jin, ‘‘A text mining and multidimensional sentiment analysis of online restaurant reviews,’’ J. Quality Assurance Hospitality Tourism, vol. 18, no. 4, pp. 465–492, Oct. 2017.

[7] R. Murphy. (Dec. 9 2020). Local Consumer Review Survey 2020. BrightLocal. Accessed: Nov. 5, 2021. [Online]

[8] R. Feldman, ‘‘Techniques and applications for sentiment analysis,’’ Commun. ACM, vol. 56, no. 4, pp. 82–89, 2013.

[9] H. Kang, S. J. Yoo, and D. Han, ‘‘Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews,’’ Expert Syst. Appl., vol. 39, no. 5, pp. 6000–6010, 2012.

[10] L. Li, L. Yang, and Y. Zeng, ‘‘Improving sentiment classification of restaurant reviews with attention-based bi-GRU neural network,’’ Symmetry, vol. 13, no. 8, p. 1517, Aug. 2021.

[11] O. Oueslati, A. I. S. Khalil, and H. Ounelli, ‘‘Sentiment analysis for helpful reviews prediction,’’ Int. J. Adv. Trends Comput. Sci. Eng., vol. 7, no. 3, pp. 34–40, Jun. 2018.

[12] E. Asani, H. Vahdat-Nejad, and J. Sadri, ‘‘Restaurant recommender system based on sentiment analysis,’’ Mach. Learn. with Appl., vol. 6, Dec. 2021, Art. no. 100114.

[13] N. M. Sharef, H. M. Zin, and S. Nadali, ‘‘Overview and future opportunities of sentiment analysis approaches for big data,’’ J. Comput. Sci., vol. 12, no. 3, pp. 153–168, Mar. 2016.

[14] B. Yu, J. Zhou, Y. Zhang, and Y. Cao, ‘‘Identifying restaurant features via sentiment analysis on yelp reviews,’’ 2017, arXiv:1709.08698.

[15] G. Beigi, X. Hu, R. Maciejewski, and H. Liu, ‘‘An overview of sentiment analysis in social media and its applications in disaster relief,’’ in Sentiment Analysis and Ontology Engineering. 2016, pp. 313–340.

[16] O. Harfoushi, D. Hasan, and R. Obiedat, ‘‘Sentiment analysis algorithms through azure machine learning: Analysis and comparison,’’ Modern Appl. Sci., vol. 12, no. 7, p. 49, Jun. 2018.

[17] B. Chopard and M. Tomassini, ‘‘Particle swarm optimization,’’ in An Introduction to Metaheuristics for Optimization. Cham, Switzerland: Springer, 2018, pp. 97–102.

[18] J. C. Bansal, ‘‘Particle swarm optimization,’’ in Evolutionary and Swarm Intelligence Algorithms. Dhahran, Saudi Arabia: Springer, 2019, pp. 11–23.

[19] S. Sengupta, S. Basak, and R. A. Peters, II, ‘‘Particle swarm optimization: A survey of historical and recent developments with hybridization perspectives,’’ Mach. Learn. Knowl. Extraction, vol. 1, no. 1, pp. 157–191, 2019.

[20] A.-Z. Ala'M, A. A. Heidari, M. Habib, H. Faris, I. Aljarah, and M. A. Hassonah, ‘‘Salp chain-based optimization of support vector machines and feature weighting for medical diagnostic information systems,’’ in Evolutionary Machine Learning Techniques. Singapore: Springer, 2020, pp. 11–34.

[21] J. Yousif and M. Al-Risi, ‘‘Part of speech tagger for Arabic text based support vector machines: A review,’’ ICTACT J. Soft Comput., vol. 9, no. 2, pp. 1–7, Jan. 2019.

[22] A. Apsemidis and S. Psarakis, ‘‘Support vector machines: A review and applications in statistical process monitoring,’’ Data Anal. Appl., Comput., Classification, Financial, Stat. Stochastic Methods, vol. 5, pp. 123–144, Apr. 2020.

[23] J. Nalepa and M. Kawulok, ‘‘Selecting training sets for support vector machines: A review,’’ Artif. Intell. Rev., vol. 52, pp. 857–900, Jan. 2019.

[24] A. Gosain and S. Sardana, ‘‘Handling class imbalance problem using oversampling techniques: A review,’’ in Proc. Int. Conf. Adv. Comput., Commun. Informat. (ICACCI), Sep. 2017, pp. 79–85.

[25] A. Fernández, S. Garcia, F. Herrera, and N. V. Chawla, ‘‘SMOTE for learning from imbalanced data: Progress and challenges, marking the 15- year anniversary,’’ J. Artif. Intell. Res., vol. 61, pp. 863–905, Apr. 2018.

[26] Obiedat, Ruba, et al. "Sentiment analysis of customers’ reviews using a hybrid evolutionary svm-based approach in an imbalanced data distribution." *IEEE Access* 10 (2022): 22260-22273.