Sentiment Analysis of Customers' Reviews Using a Hybrid Evolutionary SVM-Based Approach in an Imbalanced Data Distribution

ABSTRACT

Online media has an increasing presence on the restaurants' activities through social media websites, coinciding with an increase in customers' reviews of these restaurants. These reviews become the main source of information for both customers and decision-makers in this field. Any customer who is seeking such places will check their reviews first, which usually affect their final choice. In addition, customers' experiences can be enhanced by utilizing other customers' suggestions. Consequently, customers' reviews can influence the success of restaurant business since it is considered the final judgment of the overall quality of any restaurant. Thus, decision-makers need to analyze their customers' underlying sentiments in order to meet their expectations and improve the restaurants' services, in terms of food quality, ambiance, price range, and customer service. The number of reviews available for various products and services has dramatically increased these days and so has the need for automated methods to collect and analyze these reviews. Sentiment Analysis (SA) is a field of machine learning that helps analyze and predict the sentiments underlying these reviews. Usually, SA for customers' reviews face imbalanced datasets challenge, as the majority of these sentiments fall into supporters or resistors of the product or service. This work proposes a hybrid approach by combining the SupportVector Machine (SVM) algorithm with Particle Swarm Optimization (PSO) and different oversampling techniques to handle the imbalanced data problem. SVM is applied as a machine learning classification technique to predict the sentiments of reviews by optimizing the dataset, which contains different reviews of several restaurants in Jordan. Data were collected from Jeeran,

a well-known social network for Arabic reviews. A PSO technique is used to optimize the weights of the features, as well as four different oversampling techniques, namely, the Synthetic Minority Oversampling Technique (SMOTE), SVM-SMOTE, Adaptive Synthetic Sampling (ADASYN) and borderline-SMOTE were examined to produce an optimized dataset and solve the imbalanced problem of the dataset. This study shows that the proposed PSO-SVM approach produces the best results compared to different classification techniques in terms of accuracy, F-measure, G-mean and Area Under the Curve (AUC), for different versions of the datasets.

**EXISTING SYSTEM**

As explained by Tubishat *et al.* [26], SA, which is also referred to as opinion mining, is a text-classification field in which people's opinions, evaluations, attitudes, moods, and emotions regarding a service or product are analyzed to detect orientations. It is conducted computationally by using natural language processing, linguistics, or text analysis to detect the feelings expressed within informal text posted online. Recently, due to the popularity of social networks and online review websites, people tend to check a restaurant's reviews before visiting it. As a result, customer's impressions have become a vital factor influencing the success of restaurants; the interest among decision- makers toward customers' experience about services provided has also increased as stated by Sharif *et al.* [19]. SA has been applied to online reviews

about restaurants in the literature.

For instance, Gan *et al.* [5] studied the attributes representing consumers' reviews of restaurants. This study found that the attributes derived from previous studies such as food, service, ambiance, and price were not enough to affect restaurants' ratings and that context should be added as a significant attribute. Meanwhile, Aye and Aung [1] proposed a Myanmar language resource for lexicon-based SA as a solution to language-specific problems since most studies have considered the English language for SA. Restaurant review data were used, but informal expressions were not addressed.

Since online booking websites gained substantial interest recently, and since people now check hundreds of reviews before making any booking decisions, Agüero-Torales *et al.* [27] proposed a cloud-based software tool to analyze data from the TripAdvisor website by conducting SA on them in the province of Granada. The SA task was accomplished by using various datasets, such as the Yelp dataset, to examine the approaches proposed by other researchers as explained by Hegde *et al.* [28] and made public for research and academic studies. The Zomato Restaurant Dataset is derived from the online multinational restaurant aggregator in which reviews are posted alongside information, menus, and delivery options. Also, Taneja *et al.* [29] discussed that Zomato is a very rich database that includes information on more than 20,000 restaurants. Zomato API enables users to access the most up-to-date content and generate information about nearby restaurants. Furthermore, SemEval Datasets are high-quality annotated datasets generated through a series of international workshops; different versions of these datasets (e.g., SemEval-2015, SemEval-2016) have been used in literature as stated by Khan *et al.* [30].

Various ML algorithms have been used to conduct SA in the restaurants domain. NB, LR, and DT. ML algorithms were applied by Hassan *et al.* [32] to conduct SA on three different datasets, namely, the Yelp dataset, IMDB dataset, and Arabic qaym.com restaurant reviews dataset. Performance was measured in terms of accuracy and recall. NB and LR recorded the best results. Similarly, NB, SVM, multilayer perceptron, DT, k-NN, and fuzzy logic were applied by Kumar and Jaiswal [33] on data extracted from Twitter and Tumblr, which are widely used micro-blogging social networks. A comparative analysis of performance is presented in terms of precision, recall, and accuracy. Besides, a deep learning model called DOC-ABSADeepL was proposed by Zuheros *et al.* [34] and applied on the TripAdvisor dataset for restaurants to categorize the aspects included in an expert review while also extracting opinions and criteria. The tripR-2020 dataset was built, manually annotated, and released before being used in the same study.

**Disadvantages**

* The system is not implemented Pseudo-Code of Particle Swarm Optimizer to optimize datasets.
* In The support vector machine (SVM) algorithm is a supervised classifier that is not applied widely to solve classification and regression problems.

Proposed System

This research proposes an evolutionary approach to analyzing people's sentiments regarding restaurants' reviews in the Arabic language. Furthermore, this work followed an evolutionary hybrid approach by combining the PSO evolutionary algorithm with different oversampling techniques and the SVM algorithm to automatically detect the sentiment in the customers' comments. Four different oversampling techniques are applied to handle the problem of imbalance in

the dataset. Additionally, the applied evolutionary algorithm helps reduce the effort and time needed to tune the parameters and optimize the classification by finding the best feature weights and best *k* value for the oversampling technique, thereby resulting in better performance measures.

This work collects the reviews for almost 3000 restaurants from the Jeeran website. After the data preparation process, four different versions of the dataset are presented using different tokenization methods. The initial individual of the study is created of random weights and a random k value for the oversampling parameter. The weighted oversampled data is then classified using the SVM classification technique and the results are evaluated using G-mean. A Particle Swarm Optimizer (PSO) evolutionary algorithm is then used to optimize the values of the individual and achieve a better G-mean. Finally, the proposed approach is compared with different standard and powerful classification models, including SVM, XGBoost, DT, RF, NB, k-NN, and LR based on Accuracy, *F*1*P*, *F*1*N* , G-mean, and AUC evaluation measures.

**Advantages**

\_ Collecting the dataset from the Jeeran website with approximately 3000 restaurant reviews. The dataset is cleaned, labeled, formatted, and stemmed.

\_ Oversampling the dataset using four different oversampling techniques to solve the imbalanced problem.

\_ Applying the PSO optimization technique to find the best weights for the dataset features and the best *k* value for each oversampling technique, then applying the SVM classification technique to the oversampled and weighted dataset to find the sentiments of the restaurant reviews.

**SYSTEM REQUIREMENTS**

➢ **H/W System Configuration:-**

➢ Processor - Pentium –IV

➢ RAM - 4 GB (min)

➢ Hard Disk - 20 GB

➢ Key Board - Standard Windows Keyboard

➢ Mouse - Two or Three Button Mouse

➢ Monitor - SVGA

**SOFTWARE REQUIREMENTS:**

* **Operating system :** Windows 7 Ultimate.
* **Coding Language :** Python.
* **Front-End :** Python.
* **Back-End :** Django-ORM
* **Designing :** Html, css, javascript.
* **Data Base :** MySQL (WAMP Server).