Prediction of sentiment polarity in restaurant reviews using an ordinal regression approach based on evolutionary XGBoost

Dana A. Al-Qudah[1,2]**, Ala' M. Al-Zoubi**[2]**, Alexandra I Cristea**[3]**, J.J. Merelo**[4]**, Pedro A. Castillo**[4]**, and Hossam Faris**[1,5]

[1]**King Abdullah II School for Information Technology, The University of Jordan, Amman, Jordan**

[2]**Faculty of Information Technology, Applied Science Private University, Amman, Jordan**

[3]**Department of Computer Science, University of Durham, Durham, U.K**

[4]**Department of Computer Architecture and Technology, University of Granada, Spain**

[5]**Research Centre for Information and Communications Technologies of the University of Granada (CITIC-UGR), University of Granada, Granada, Spain**

Corresponding author:

Hossam Faris[15]

Email address: hossam.faris@ju.edu.jo

ABSTRACT

As the business world shifts to the web and tremendous amounts of data become available on multilingual mobile applications, new business and research challenges and opportunities have been explored. This research aims to intensify the usage of data analytics, machine learning, and sentiment analysis of textual data to classify customers' reviews, feedback, and ratings of businesses in Jordan's food and restaurant industry. The main methods used in this research were sentiment polarity (to address the challenges posed by businesses to automatically apply text analysis) and bio-metric techniques (to systematically identify users' emotional states, so reviews can be thoroughly understood). The research was extended to deal with reviews in Arabic, dialectic Arabic, and English, with the main focus on the Arabic language, as the application examined (Talabat) is based in Jordan. Arabic and English reviews were collected from the application, and a new model was proposed to sentimentally analyze reviews. The proposed model has four main stages: data collection, data preparation, model building, and model evaluation. The main purpose of this research is to study the problem expressed above using a model of ordinal regression to overcome issues related to misclassification. Additionally, an automatic multi-language prediction approach for online restaurant reviews was proposed by combining the eXtreme Gradient Boosting (XGB) and particle swarm optimization (PSO) techniques for the ordinal regression of these reviews. The proposed PSO-XGB algorithm showed superior results when compared to support vector machine and other optimization methods in terms of root mean square error (RMSE) for the English and Arabic datasets. Specifically, for the Arabic dataset, PSO-XGB achieved an RMSE value of 0.7722, whereas PSO-SVM achieved an RSME value of 0.9988.

INTRODUCTION

Social online eating, food ordering, and related applications and websites have become common in people's lives (*Roh and Park*, 2019). However, while the food industry is growing, many businesses are closing down (*Martín-Valdivia et al.*, 2013; *Kauer and Moreira*, 2016). Restaurants and their reputations and customers' feedback (*Vinodhini and Chandrasekaran*, 2012). However, receiving useful feedback can be challenging; hence, online ordering apps and websites usually allow customers to directly provide feedback and ratings. This way, business owners can be frequently updated on their customers' needs and opinions and make enhancements to their services accordingly. However, since manually tracking hundreds or thousands of feedback messages is almost impossible, feedback tracking remains challenging.

Currently, sentiment polarity (SP) is the recommended way to address this issue. SP is the process of dealing with written text in order to identify and interpret the opinions expressed in it. These opinions are usually categorized, for example, as positive, negative, or neutral (*Krishna et al.*, 2019; *AlZu'bi et al.*, 2022). The need to use SP is derived from the tremendous amount of data collected, extracted, loaded, and used in structured and unstructured text harvested from the internet (*Ravi and Ravi*, 2015). SP uses combined or separate natural language processing (NLP) approaches and various text analysis techniques so that the reviews and the opinions expressed therein can be classified based on various criteria (*Krishna et al.*, 2019).

However, there is no single best procedure for applying SP, as it is a wide, developing area of computational text analysis. Researchers have used different performance evaluation methods, such as text classification, including naïve Bayes, decision tree classifiers, and N-fold cross-validation (*Karsi et al.*, 2017; *Adnan et al.*, 2019). There are two main sentiment polarity approaches: knowledge-based or machine learning-based SP. The main difference between these two approaches is that the former is based on NLP algorithms or lexicon methods (*Neethu and Rajasree*, 2013), whereas the latter focuses on data polarity (i.e., whether a text is negative, positive, or neutral) based on training on data previously labeled by humans (*Gautam and Yadav*, 2014).

Arabic NLP is a relatively new research area, especially for Dialectal Arabic. One such dialect that is particularly underexplored is the Jordanian dialect. Moreover, as Arab speakers are often multilingual and use more than one language in their daily lives, a multi-language approach is required. However, very few studies of this nature exist (*Dashtipour et al.*, 2016). Talabat, a multilingual restaurant and food ordering application, was examined in this study to explore users' reviews, feedback, and ratings of restaurants in Jordan.

As mentioned above, the dramatic expansion of application usage and the enormous availability of text-based data have raised many challenges for business owners and application developers. The main challenges are related to the need to analyze customer reviews and categorize them as positive, negative, or neutral. Each polarity helps business owners maintain their advantages and improve their shortcomings in different ways. Also, rating services require combined analysis with text reviews, as ratings alone do not give enough information for business owners to understand them. The sentiment underlying each review must be clear, and automated procedures should be provided for business owners to understand how reviews and ratings are labeled. These challenges force developers and researchers to find enhanced methods for searching for, learning about, and evaluating the sentiments and ratings of reviewers in both Arabic and English languages.

The misclassification problem between multi-class cases was addressed using a model based on ordinal regression formulation in this research. Ordinal regression can help determine the link between distinct classes, which is difficult to determine using other methods. This model was also proposed to cover all the matters discussed above. The model was construed as follows. The first phase was data description and collection, during which the dataset was built based on Arabic and English reviews and ratings collected from the Talabat application. In the next step (data preparation), processes such as stemming, cleansing, and rooting were conducted in both Arabic and English. The proposed approach was then explained thoroughly, the model was evaluated, and the main results were explained and interpreted. This work addresses two main research questions:

- How can sentiment analysis methods and techniques improve the quality of feedback extracted from reviews and ratings in Arabic (with its dialects) and English?

- Can the proposed method optimize the sentiments of feedback in Arabic and English compared to state-of-the-art methods?

In previous work, (*Al-Qudah et al.*, 2020) demonstrated the superiority of the extreme gradient boosting (XGBoost) algorithm in classic classification problems for sentiment analysis. The algorithm effectively predicted and analyzed customers' opinions of an e-payment service when the researchers combined a neutrality detector model with XGBoost and a genetic algorithm to solve a classic multi-class problem. However, in the current study, XGBoost is applied to entirely different problems, namely ordinal regression in sentiment polarity. This research aims to address the challenges posed by businesses by analyzing online restaurant reviews in a multi-language environment.

The main contribution of this research is the proposal of an ordinal regression formulation to minimize the misclassification gap between multi-class problems. Moreover, it combines the swarm-intelligent

optimizer with the powerful XGBoost algorithm to handle multi-class sentiment classification problems. The proposed approach has two primary advantages. Firstly, the swarm-intelligent optimizer automatically tunes the parameters of the XGBoost algorithm, eliminating the need for human intervention to complete this task. Secondly, it handles the classification problem as an ordinal multi-class classification problem, considering the importance of class order. One of the main features of the proposed data harvesting and cleaning approach is that feedback messages in both Arabic and English were harvested. Dialectical Arabic was also considered due to the lack of standardization for dialects and their sheer number. After data cleaning was performed for the dataset, a particle swarm optimization (PSO) algorithm was used to optimize the XGBoost algorithm and detect the five classes of sentiment polarity. Ordinal regression was also used to classify the data.

The questions answered in this study are as follows: How effective are current sentiment analysis methods in accurately identifying the sentiments in reviews? Does incorporating multi-language support, specifically for the Arabic language, improve the accuracy of sentiment analysis? Can the proposed method using ordinal regression and PSO-SVM with the XGBoost algorithm increase the accuracy of sentiment classification? What specific challenges are faced in processing and analyzing sentiments in dialectal Arabic?

The main aims of the study are as follows:

- To study customers' reviews, feedback, and ratings in Arabic in general and the dialectal Arabic used in Jordan using a popular restaurant ordering and reviewing application called Talabat.

- To address the challenges faced by restaurant owners and application developers in effectively organizing and comprehending large amounts of available text-based data as feedback for their businesses.

- To delve into customers' feedback, ratings, and reviews, not as plain text data but in a way that enables the classification of the sentiments of feedback into positive, negative, and neutral categories .

- To conduct a comprehensive analysis amalgamating the reviews and feedback with rating services, mixing two different types of data analysis—namely, text data and categorical data—to provide a holistic understanding of customer perceptions.

The remainder of the paper is organized as follows. First, previous studies on ordinal regression sentiment polarity are introduced in the 'Related Work' section. Then, the methods used in this study are described in the 'Preliminaries' section. Next, the proposed approach is discussed in the 'Methodology' section. The results of the experiments are subsequently analyzed in the 'Experiment and Results' section. Finally, the conclusions and future directions are presented in the 'Conclusion' section.

RELATED WORK

The food service industry is economically important, and many of its services have moved online. Therefore, much research has been conducted to improve the specific services provided to online users to make ordering and receiving food using these services easier and more satisfying. This research analyzes the text harvested from users' feedback on such systems. Many researchers have conducted similar work that has enriched the state of the art. However, this work presents a new combination of algorithms that has not been sufficiently explored while considering both the Arabic and English languages in the Jordanian market specifically.

Sun et al. (2019) recently conducted a study on recommender systems of online Chinese restaurants following uncertainty theory and using sentiment polarity. They highlighted the importance of analyzing customer reviews, as the overall ratings did not accurately represent their reflections and opinions. Furthermore, they claimed that these reviews should be divided based on opinions' similarities and the designs of the utilized recommender systems. They analyzed the text provided by users by determining the main attributes of the sentiment polarity. This was done by acquiring the main attributes from the reviews by performing a fine-grained classification, which eventually aids in assessing the polarity and strength of text. This part of the study was conducted using HowNet (*Dong and Dong*, 2006). The researchers continued their work by using the uncertainty theory to build the recommender algorithms

and models. Finally, they used KNN and K-means algorithms to further classify and cluster reviews while also determining the accuracy of these recommender systems.

Adnan et al. (2019) analyzed restaurant reviews using a decision tree (J48) algorithm. They used users' reviews of and comments about restaurants in Surabaya on TripAdvisor and harvested the data using web scraping software called WebHarvy, which saves data in Excel format. The data contained information such as the name of the customer; their rating, comment text, and comment title; and the name of the restaurant. The researchers also pre-processed the data using the Natural Language Toolkit available in Python. In the pre-processing stage, many operations were done, such as tokenization, slang word removal, stop word removal, and symbol removal. After the cleaning and data processing stages, the researchers calculated the number of appearances of each word in the text using a J48 decision tree. They obtained a precision of 48, a recall of 36.8, an accuracy of 45.6, and an f-measure of 41.4, indicating a classification recommendation to identify good restaurants.

Krishna et al. (2019) conducted more comprehensive research by studying various machine learning techniques to analyze the sentiments of restaurant reviews. They used a dataset in Tab Spaced Values format. They pre-processed the data by applying various cleaning and tokenization techniques. They then prepared bags of words created from pairs of different documents by performing disjoint unions of the words and summing up their multiplicities. Afterwards, they applied various classification algorithms, such as naïve Bayes, support vector machine (SVM), decision tree, and random forest. The experiment was run three times, and SVM had the highest prediction accuracy (94.56).

Zhang et al. (2015) focused on the type of sentiment analysis conducted, namely, whether the analysis is on a document-review level, a structure level, or a phased level. The results suggest that precision can reach 90% when comparing sentiment analysis results from document level or structure level. Meanwhile, precision can be around 70–80% when using phrase-level analysis.

From another perspective, sentiment polarity in text has also been explored using XGBoost, with researchers mainly analyzing tweets. This is because XGBoost performs well when applied to large-scale problems, as it employs highly flexible operations and calculations on most regression, classification, and ranking problems (*Chen and Guestrin*, 2016). XGBoost has been used in many studies related to sentiment polarity, such as the work of (*Jabreel and Moreno*, 2018). They used XGBoost to analyze text extracted from tweets, focusing on lexicons-based features. They compared this approach with a deep learning approach called N-Stream ConvNets. The main outcome of their research is that combining an ensemble technique with XGBoost helped improve the performance of the booster. The researchers suggested that this combination provides similar or better results compared to those given by the deep learning technique.

Furthermore, sentiment polarity was discussed by *Kern et al.* (2021), who studied sentiment analysis in German language using cluster analysis. The researchers used an unsupervised method to cluster the polarity of the German language, with the results showing that different German dictionaries were similar; the main differences were detected in the structure and outliers. They based their conclusions on a k=3 cluster using k-means.

Meanwhile, *Bhoi and Joshi* (2018) discussed how to classify a certain aspect in a sentence. They argued that sentiment polarity should be used for classification not by looking at a sentence as a bag of words but rather at items in temporal order. The authors used common pre-processing steps before testing many algorithms, such as naïve Bayes, decision trees, support vector machine, random forest classifier, extra trees classifier, and XGBoost, combined with other deep learning techniques. Their results indicate that XGBoost performed well as a classifier and that other deep-learning algorithms could provide more accurate results.

Nobre and Neves (2019) studied sentiment in the financial market by combining principle component analysis (PCA), XGBoost, genetic algorithms, and discrete wavelet transform (DWT). PCA and DWT were combined into one system to accomplish high returns while minimizing possible risks. This combined approach yielded good results, with an average rate of return of 49.26 in the portfolio. In other work, *Song et al.* (2020) explored steel property optimization using both PSO and XGBoost. The authors attempted to build an optimization model with 27 features for machine learning to predict the tensile strength and plasticity. The complexity of the model alongside the numerous features was studied using the combined approach. The main results indicate that XGBoost was the most reliable prediction model examined in the study. The PSO approach aided the final optimization of the model for iron and steel production.

Le et al. (2019) combined the PSO and XGBoost methods and found that smart cities play an important role in the development of countries. They recorded higher reliability for their combined method than for other machine learning algorithms regarding the prediction and optimization of the heating loads of buildings. Although the work presented by these researchers targeted different applications, they proposed the combination of PSO and XGBoost, which the current work further explores in a different domain, in addition to ordinal regression algorithms, as explained further.

Huang et al. (2019) explored sentiment analysis on social media data using a novel approach. This approach primarily revolves around the integration of three distinct attention models for predicting sentiments. Their experiments showcase the efficacy of utilizing various types of datasets, including weekly labeled data and manually labeled datasets, alongside three different models tailored for sentiment classification. Notably, their methodology employed deep multi-modal fusion to discern features and establish correlations between visual and textual content. They employed a mixed fusion framework to merge sentiment analysis, emphasizing regional word features to differentiate between models. Their proposal encompassed models designed to effectively learn emotion classifiers for both visual and textual content.

Huang et al. (2020) work extends previous discussions by introducing a novel model known as the attention-based modularity gated network. This model determines correlations between image and text modalities while extracting discriminative features for multi-modal sentiment analysis. Specifically, the authors employed a visual semantic attention model to learn visually attended features for each word, effectively integrating sentiment information from both modalities. Additionally, they propose a long short-term memory (LSTM) network to adaptively learn multimodal features, selecting modalities that exhibit stronger sentiment signals. The model incorporates a self-attention mechanism to enhance semantic understanding.

Later work delved into multi-modal approaches for analyzing both visual and textual data (*Thuseethan et al.*, 2020). This study highlights the potential pitfalls of blindly merging textual and image data for sentiment analysis and classification and proposes that the interrelations between multi-modal data should be explored using a deep association learner, which is adept at discerning relationships by leveraging learned visual and textual features, thereby automatically discriminating between features extracted from text and images. The two distinct streams of model features can then be extracted, focusing on the most pertinent aspects related to sentiment. Subsequently, sentiment estimation was performed through a late fusion mechanism. Comprehensive evaluations demonstrate promising results, indicating the capability to classify data based on sentiments, whether they comprise text, images, or a combination of both.

(*Naeem et al.*, 2022)implemented various machine learning models to gauge sentiment polarity in user reviews on the Internet Movie Database. The process involved preprocessing the reviews to eliminate noise and redundant information, followed by employing classification models such as support vector machines (SVMs), naïve Bayes, random forest, and gradient boosting. Feature engineering techniques, including TF-IDF, a bag of words, global vectors for word representations, and Word2Vec were applied, alongside hyperparameter tuning, to enhance classification accuracy. The results reveal that SVM, when combined with TF-IDF features, achieved the highest accuracy of 89.55%. However, user sentiment contradictions pose a challenge to accurate classification. To address this, TextBlob was employed to assign sentiment labels to the review dataset. The results of TextBlob-assigned sentiments indicate a potential accuracy of 92% when using the proposed model.

Another study showed that sentiment expressed in tweets regarding deep fakes holds considerable importance in understanding public perception (*Rupapara et al.*, 2021). This study introduced a deep learning approach to assess the sentiment polarity of such tweets and proposed a stacked bi-directional long short-term memory (SBi-LSTM) network for sentiment classification. Additionally, various traditional machine learning classifiers, such as support vector machine, logistic regression, Gaussian naive Bayes, extra tree classifier, and AdaBoost classifier, were explored alongside feature extraction methods like term frequency-inverse document frequency, and bag of words. The performances of deep learning models, including long short-term memory network, gated recurrent unit, bi-directional LSTM, and convolutional neural network+LSTM, were also evaluated. The findings demonstrate that the SBi-LSTM model outperformed both traditional machine learning and deep learning approaches, achieving an accuracy of 92%.

Rupapara et al. (2021) conducted another study in the same context by detecting fake news, which has become a crucial area of research, particularly in languages like Urdu, which is spoken by over 230 million

people, but for which investigations remain limited. This study evaluated the effectiveness of various machine learning classifiers and a deep learning model in detecting fake news in Urdu. Classifiers such as logistic regression, support vector machine, random forest, naïve Bayes, gradient boosting, and passive aggression were employed, alongside analysis of term frequency-inverse document frequency and bag of words features. Based on a dataset of 900 manually collected news articles, the study found that random forest performed the best, achieving an accuracy of 0.92 with bag of words features. In comparison, machine learning models generally outperformed deep learning models such as long short-term memory and multi-layer perceptron in this context.

The research also discussed that as restaurants have joined online platforms like UberEATS, Menulog, or Deliveroo, customer reviews have become vital for assessing company performance (*Adak et al.*, 2022). Food delivery service (FDS) organizations seek to utilize customer feedback to address complaints and enhance customer satisfaction. This study examines machine learning (ML), deep learning (DL), and explainable artificial intelligence (XAI) methods for predicting customer sentiments in the FDS sector. A review of the existing literature highlights the prevalence of lexicon-based and ML techniques for sentiment prediction based on customer reviews, with a limited adoption of DL due to interpretability concerns. Key findings reveal that many models lack interpretability, posing challenges for organizational trust. While DL models offer high accuracy, they lack explainability. However, this issue can be addressed using XAI techniques. Future research should focus on integrating DL models into FDS sentiment analysis and incorporating XAI methods to enhance model explainability.

Shahi et al. (2022) focused on Nepali-language COVID-19-related tweets and proposed an analysis of sentiments using both syntactical and semantic information. They combined TF-IDF and FastText text representation methods to create hybrid features for enhanced discrimination. Nine widely used machine learning classifiers were implemented based on three feature rep- representation methods: TF-IDF, FastText, and a hybrid method. The methods were evaluated using a NepCov19Tweets dataset, with data categorized into positive, negative, and neutral classes. The results indicate that the hybrid feature extraction method outperformed individual methods across nine machine learning algorithms, demonstrating superior performance compared to state-of-the-art techniques.

Ordinal regression has been shown to be effective when working with labeled and unlabeled data. *Rafique et al.* (2022) focused on understanding the accuracy of transductive ordinal regression as a highly accurate algorithm when applied to pre-processed data. This type of algorithm is well-developed and is proposed to work prominently with sentimentally labeled data in the Bulgarian language. In addition, *Kapukaranov and Nakov* (2015) studied movie reviews in Bulgarian using ordinal regression. They used multiple classification algorithms to set a threshold region to classify positive, negative, and neutral reviews. The aim was to predict which of these regions would be set to be positive, negative, and neutral. They divided the regions based on the sentiments of the reviews.

Meanwhile, *Loke et al.* (2020)) studied sentiment polarity using neural network architecture. They studied attention-based sentiment analysis and used neural networks to classify the outcomes. They used F1 scores, accuracy loss function, and ROC AUC to evaluate the results. scores. The results indicate F1 scores of around 72%, accuracy of 92%, and losses of around 20.

Moreover, *Saad and Yang* (2019) examined the sentiment polarity of Twitter data (tweets) using different machine learning algorithms to deal with ordinal regression problems. The results indicate highly accurate prediction rates in comparison to other algorithms such as SVM, RF, and DT. The researchers combined the ordinal regression classifier with XGBoost and PSO.

Al-Qudah et al. (2020) employed XGBoost with the genetic algorithm for parameter optimization in a classification problem. The researchers predicted the sentiments (positive, negative, or neutral) of e-payment service reviews in the Arabic language. Accordingly, in the present paper, XGBoost was implemented in order to handle ordinal regression in sentiment polarity (ratings of 1, 2, 3, 4, and 5) for a multi-language problem. Furthermore, PSO was used to optimize the XGBoost parameters, thereby enhancing its performance. Unlike the previous work, this study handled ordinal regression for the sentiment polarity issue while investigating two different languages.

This approach differs from the previous works by combining PSO and XGBoost to optimize parameters for the ordinal regression problem for multilingual restaurant reviews. The selected languages involve different pre-processing techniques and require more analysis to deal with . Furthermore, XGBoost was compared with various metaheuristic algorithms, including PSO and the well-known whale optimization algorithm (WOA). *Mirjalili and Lewis* (2016) and multi-verse optimization (MVO) (*Mirjalili et al.*,

2016). Then, the best algorithm was chosen for another comparison with support vector machine, another well-known classifier.

Based on the extensive study of the recent literature and case studies, a comprehensive understanding of the existing work in the field has been obtained. The main contributions of our work, in comparison to the available studies, are categorized into three main domains: data source, modeling techniques, and optimization. Most of the data used in previous studies were collected from social media—for example, Twitter (now known as X) and Facebook. Moreover, in one case, the data source was a delivery service platform. Our data source is one of the most important and widely used food ordering applications, whose business model includes reviews of the delivery service model. In addition, this study examines the usage of ordinal regression to overcome issues related to misclassification. Such issues have not been thoroughly discussed in the literature as they relate to the Arabic language or while considering the dialectical Arabic used in Jordan. The final contribution of this study relates to the optimization of the utilized models; few papers have discussed optimization in sentimental analysis. Meanwhile, our research utilized two main optimizers.

PRELIMINARIES

Ordinal regression/classification

Ordinal classification or regression is a type of multiclass classification method for classes that have an ordering relationship, even though they do not have any meaningful numeric differences (*Gaudette and Japkowicz*, 2009). This problem is considered one of the most significant tasks in relation-learning. Ordinal data are frequently classified in scenarios involving human-made scale problems. In other words, classes may contain different sizes of values such as small, medium, and large (*Yıldırım et al.*, 2019) or cheap, normal, and expensive. Meanwhile, ordinal categorical variables can be factors or predictors in several statistical procedures, like linear regression.

The purpose of ordinal data prediction is to determine how to calculate distances between categories without knowing how to calculate distances between variables. These issues lay between the categorization and regression techniques utilized in psychology, sociology, and other disciplines. The modeling of human preference levels (e.g., on a scale from 1–5, depending on how strong one's preference is) is an example of ordinal regression *Bürkner and Vuorre* (2019). One of the advantages of ordinal measurement is that it simplifies the processes of collecting and categorizing data.

Class labels convey information about the order of classes; an average class vector has a higher (or better) rating than a poor class, but a good class exceeds both. Two factors are significant in this type of problem. First, different types of misclassification have various consequences; for example, misclassifying an excellent teacher by erroneously categorizing him as poor is more severe than incorrectly classifying him as very good. Second, more accurate models may be built by using ordering information.

Ordinal classification has been applied to several real-world problems, including problems related to medical sciences (*Cardoso et al.*, 2005), collaborative filtering (*Shashua and Levin*, 2003), information retrieval (*Herbrich et al.*, 1999), econometric modeling (*Mathieson*, 1996).

XGBoost algorithm

The XGBoost algorithm is an enhanced version of the gradient boosting algorithm that applies decision trees. XGBoost was developed by *Chen and Guestrin* (2016) to solve regression and classification problems efficiently and rapidly. It was developed to improve machines' speed and exploit the full functionality of their resources, such as memory and hardware. The XGBoost algorithm is valued owing to its ability to reduce time consumption while handling, problems such as missing values, parallel execution, and the use of the optimal machine resources. Furthermore, the algorithm adopts the regularization technique to decrease and prevent the overfitting problem. Further extensions, including stochastic and gradient boosting, further improve its performance (*Song et al.*, 2020).

As the base learners are tree algorithms, the XGBoost algorithm divides the dataset attributes into conditional nodes consisting of several branches and a leaf node (*Chen et al.*, 2020). Moreover, the hyperparameters of the XGBoost algorithm are considered quite important in seeking the optimal results for a specific problem; therefore, tuning these parameters is necessary.

PSO algorithm

Particle swarm optimization (PSO) is a well-known metaheuristic algorithm inspired by the behavior of flock birds (*Kennedy and Eberhart*, 1995). The population-based algorithm simulates the position of these birds in order to achieve an optimal solution. In PSO, particles are denoted as a set of solutions called a population, and the solution consists of different parameters found in a given multidimensional space. These particles (i.e., the population) are grouped to perform a swarm that searches the space at a specified velocity to find the optimal solution.

Particles can save their memory to keep tracking the former best position (solution). The positions of these particles can be adjusted until the optimal solution is discovered according to the personal best experience (pbest) and global best, which represents other members' best experiences (gbest) (*Ding et al.*, 2020). Moreover, the historical behavior of particles and their neighbors helps update the particles' velocities while searching the search space (flying). Therefore, the search process tends to improve in every iteration (*Ghamisi and Benediktsson*, 2014).

In this work, PSO was combined with XGBoost to optimize the ordinal classification of restaurant reviews by finding the optimal parameters for XGBoost.

METHODOLOGY

This section presents the methodology for detecting restaurant reviews. Four phases are described in detail: data description and collection, data preparation, the proposed approach, and model evaluation. In the first phase, the data information, the source of the data, the method used for the collection process, and the statistics details are described. In the second phase, the data are prepared. This phase consists of formatting, cleaning, stemming, and feature extraction. The third phase presents design issues, the fitness function, and the system architecture of the proposed approach. The last phase is the evaluation phase.

As for business applications, using reviews of customers to evaluate and enhance the service provided is vital. Thus, extra focus is given to studying user satisfaction using data mining algorithms, namely XGBoost combined with PSO for ordinal regression.

Data description and collection

The research targets a specialized food ordering application called Talabat (*Talabat*, 2004), which is well-known in the Middle East. It operates in countries such as Saudi Arabia, Oman, Kuwait, Qatar, the UAE, and Jordan. It works as a mediator between registered restaurants and customers. This research analyzed 2000 reviews annotated with their ratings from consumers who used Talabat. The ratings are divided into five classes, from 1 to 5, with 1 being the lowest rating and 5 being the highest. All opinions associated with the ratings were harvested using a customized Python script.

Services cannot be improved without the appropriate feedback from stakeholders, regular customers, or both. Therefore, such reviews, if acted upon, can make a service better in different ways that owners—including restaurant owners—may have neglected.

Problems can be avoided by analyzing the customers' reviews, as well as the ratings they use, which express their sentiments. Automated procedures, which save time and money, are enticing even for small businesses. Therefore, a polarity analysis is performed in the form of a rating prediction on bi-lingual data.

The reviews in this paper were collected from the Talabat website using a crawler tool. Each review contains a customer username, the date of the review, and the associated rating. The ratings were crawled as images of 1, 2, 3, 4, and 5 stars. The reviews were written in two languages: Arabic and English.

The data consist of 1927 instances and 1292 features (terms). More details on the dataset can be found in Table 1.

Table 1. Bilingual dataset description.

Dataset	Instances	Features	Classes
English	935	385	5
Arabic	992	707	5

Data preparation

In this stage, the reviews (data) went through several standard pre-processing procedures (*Faris et al.*, 2017; *Habib et al.*, 2018). In other words, these procedures prepared the data so it could be read by the classifiers. The procedures included formatting the data, removing missing values, and cleaning (*Hassonah et al.*, 2020).

The labeling process was not needed for customer reviews since the reviews had already been labeled by the customers. However, to enhance accuracy, several experts were asked to read a sample of the data and ensure that the labeling was correct.

The first step in the data preparation process was to split the data into two separate datasets based on language (English and Arabic). This had to be done due to the differences in these languages' characteristics, namely in terms of stop words, stemming processes, prefixes, and suffixes.

For the Arabic dataset, various stop words were removed so that the true meanings of sentences could be identified. Moreover, a normalization technique was applied to discard special characters and non-Arabic letters to reduce the number of extracted terms. Finally, a stemming process was performed in order to decrease the duplication of extraction terms (prefixes and suffixes). After different stemming methods were tested , the Arabic light stemmer method was chosen for the Arabic language (*Al Ameed et al.*, 2005).

Meanwhile, for the English dataset, stop words such as "the," "and," and "but" were removed. Then, special characters and symbols were removed to decrease the number of terms that could be extracted. Furthermore, the snowball stemmer was used to remove the suffixes and prefixes of the text (*Porter*, 2001).

After the previous steps were completed, a tokenization procedure was applied to break the words into tokens by analyzing the text linguistically and splitting the words. The term frequency-inverse document frequency ($TF - IDF$) was used to extract features (Ghag and Shah, 2014). This technique calculates the relevance numerically for a particular document, as shown in Equations 1 & 2. $TF(t,d)$ is the number of words (t) in the document (d) and $IDF(t)$ is the inverse document frequency.

$$IDF(t) = log(\frac{N}{DF(t)}) \tag{1}$$

The $TF - IDF$ value can be computed as follows.

$$TF.IDF(d,t) = TF(d,t) \times IDF(t) \tag{2}$$

In Equation 1, N denotes the number of all documents and $DF(t)$ is the number of documents in which the words (t) occur.

Eventually, the tokenization process (TF-IDF) was followed to convert the data into a matrix in order to run it through the classifier. However, some problems, including missing values and noisy data, could have been encountered and needed to be resolved before the data were examined. These problems can be solved by majority vote (*Xia et al.*, 2017) and normalization (*Wang et al.*, 2006).

Proposed approach

This subsection describes the proposed approach applied to the datasets. Furthermore, the sentiment identification model is assessed via a root mean square error (*RMSE*) and mean absolute error (*MAE*). The model is displayed in Fig. 1 below.

After the datasets were prepared, the classification model was run on the data. The XGBoost algorithm was the main classifier used and was compared with various state-of-the-art algorithms for the task of sentiment polarity prediction. Recently, XGBoost has gained attention in the literature, especially in machine learning applications (*Al-Qudah et al.*, 2020). Nevertheless, because it has many parameters, it is challenging to find the appropriate combinations between applications and associated problems (*Jiang et al.*, 2019).

Thus, particle swarm optimization (PSO) was applied, as it can help find the optimal combination of hyper-parameters for XGBoost. It is also faster and more accurate than the other methods described in the literature (*Lin et al.*, 2008), such as grid search. The grid search algorithm can search for the best parameters in a given range for different models. However, it takes a substantial amount of time to run and find the local optimum solution. Using PSO is more efficient for such problems. Meta-heuristic

457 algorithms (*Al-Zoubi et al.*, 2019), like PSO, have three main components: searching, learning, and
458 evaluation.

459   XGBoost has many hyper-parameters. Those that are the most frequently used in the literature were
460 selected (Table 2.

Table 2. The XGBoost parameters used and their descriptions, ranges, and best values *Al-Qudah et al.*
(2020); *Jiang et al.* (2019)

#	Parameters	Description	Range
1	min_child_weight	A leaf's minimum weight	[1-20]
2	gamma γ	Reduction minimum loss required for the partition	[0.1-5.0]
3	subsample	Ratio of training records	[0.1-1.0]
4	colsample_bytree	Features sub-sample ratio	[0.1-1.0]
5	max_depth	Tree depth maximum	[1-20]
6	learning_rate	Step size	0.02
7	n_estimators	Number of trees employed	1000

461 ### *Design issues*

462 Two design issues should be considered when an optimization algorithm is applied to a problem: the
463 design representation for the solution and the fitness function (Fig. 1).
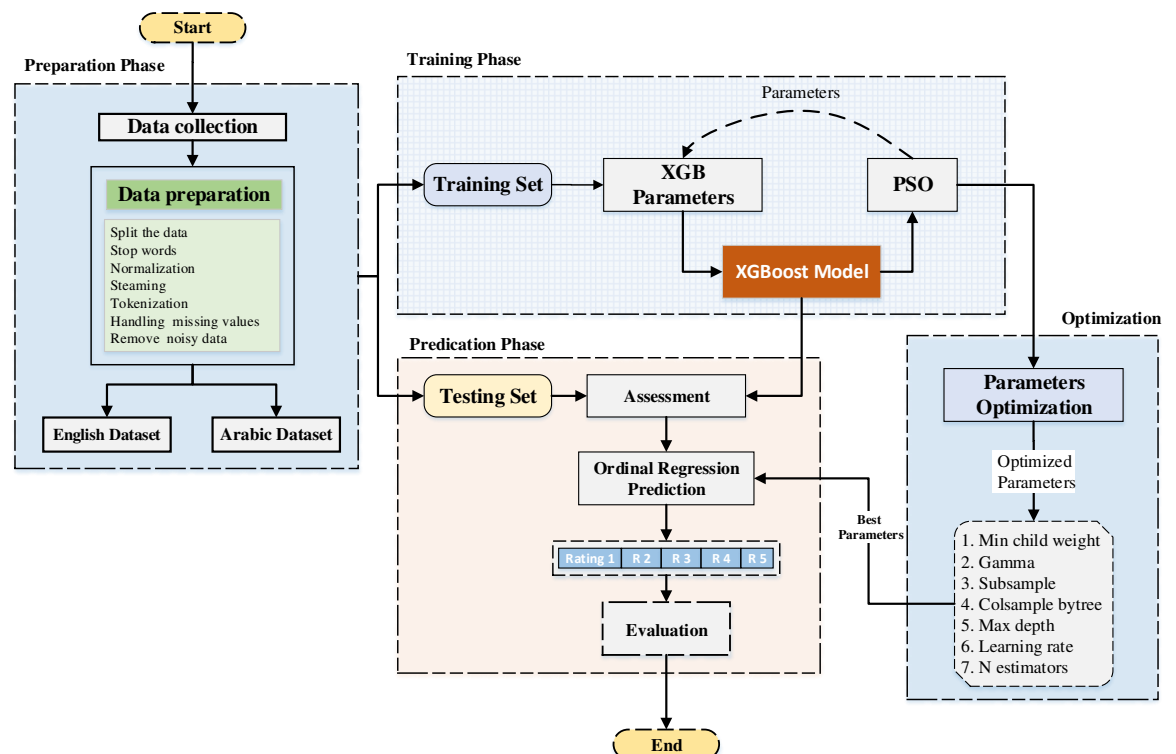
Figure 1. General overview of the applied framework

464   **Solution representation:** The PSO swarms are designed to represent the solution to the problem. In
465 this work, the main problem relates to finding the optimal parameters for XGBoost. The swarms consist
466 of a one-dimensional set of randomly generated numbers corresponding to the parameter's value. These
467 generated numbers are scaled from 0 to 1 to simplify the selection criteria of the parameters, as shown in
468 the following equation:

$$B = \frac{A - min_A}{max_A - min_A}(max_B - min_B) + min_B \qquad (3)$$

where *A* denotes the value that needs to be scaled, *B* represents the new scaled value, min_A is the lower bound, and max_A is the upper bound of the old range. The lower and upper bounds of the new range are denoted by min_B and max_B, respectively.

Fitness function: An evaluation criteria was applied to improve the generated solutions from PSO, and feedback from XGBoost was provided for every iteration. The root mean square error (*RMSE*) was selected as a fitness function. *RMSE* is the most common metric used for ordinal regression problems in the literature (*Li et al.*, 2019; *Gaudette and Japkowicz*, 2009; *Shi et al.*, 2018) due to its ability to show the degree of deviation between predicted and original labels. Therefore, the PSO algorithm was modified to minimize the fitness value (fitness function). The equation used to calculate *RMSE* is given below:

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \qquad (4)$$

where *n* is the total number of samples y_i is the actual value, and $\hat{y}_i$ is the estimated value.

System architecture: In this phase, the datasets were split into a training and a testing set using the 10-fold splitting criteria (*Shao et al.*, 2013; *Basiri et al.*, 2008). The dataset was divided into k parts, with the training set containing $k - (1/k)$ parts and the test set containing the remaining $(1/k)$ parts. This procedure ensured that the training and testing sets were differentiated and that the optimal model was attained (*Hassonah et al.*, 2020), which is useful, especially in cases when the training data is limited.

In the first iteration, PSO generated a random set of real numbers in a vector form. Then, XGBoost started the training process using the parameters selected by PSO. After the training was completed, XGBoost sent the fitness value to PSO. These steps were then repeated until the termination criteria was reached—in this case, the maximum number of iterations. Consequently, the best-selected values generated by PSO were used in the testing phase. All previous steps were repeated k times, and the average value was recorded.

Evaluation

Several evaluation measures were applied to assess and calculate the model's performance. In addition to the root mean square error (*RMSE*), the mean absolute error (*MAE*) measure was been used; it was calculated using the following equation:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \qquad (5)$$

where *n* is the total number of samples, y_i is the actual value, and $\hat{y}_i$ denotes the estimated output value.

Furthermore, an extended analyzed evaluation process was applied in order to state the errors in the predicted ordinal classes by using a confusion matrix table.

The classes that were correctly predicted were labeled as "true positive" (TP), while "false negative" (FN) was used to refer to classes that were wrongly predicted as incorrect ratings. Similarly, "false positive" (FP) represents classes that should have been predicted as wrong ratings but were predicted as correct ratings. Finally, "true negative" (TN) denotes classes that were correctly predicted as wrong ratings.

EXPERIMENTATION AND RESULTS

This section describes the results of several experiments applied to the prepared datasets. The results show that the combination of XGBoost with a metaheuristic (PSO) obtained better results than each method independently. Both datasets (Arabic and English) were processed in the same stages.

First, classical machine learning methods (as explained below) were compared with the default XGBoost algorithm. Next, in the second phase, the proposed PSO-XGB was compared to other meta-heuristic algorithms. In the final phase, the (PSO-XGB) and a classification model combined with the best metaheuristic were compared. Two evaluation measures were applied to these models, namely, root mean square error (*RMSE*) and mean square error (*MAE*).

All experiments were conducted on an Intel Core i5-6400 personal computer with 8GB RAM. The proposed model was implemented on Python 3.7.

515 First, XGBoost was applied to the English and Arabic datasets and compared with classic machine
516 learning models (J48, RF, k-NN, and NB). Then, PSO and other metaheuristic algorithms (WOA and
517 MVO) were combined with XGBoost and compared. Furthermore, a feature importance analysis was
518 carried out. Finally, PSO was combined with another classification model (SVM) (*Ala'M et al.*, 2020)
519 and compared with the proposed PSO-XGB model. Moreover, a detailed examination was conducted to
520 state the errors of the model using a confusion matrix. All experiments adhered to the 10-fold criteria.

521 **Phase 1: Comparison with classic machine learning models**
522 In this phase, the default XGBoost algorithm was applied to the English dataset and compared to classic
523 machine learning models.
524 As shown in Table 3, XGBoost achieved the best RMSE result of 1.2472. The next-best result was
525 obtained by RF (2.7373). As for the MAE measure, NB yielded the best result of 0.9613, followed by
526 XGB (0.9925).

Table 3. Resuls for the English dataset

Algorithm	*RMSE*	*MAE*
J48	3.0384	1.4916
RF	2.7373	1.1362
k-NN	2.9898	1.0547
NB	2.9725	0.9613
XGB	1.2472	0.9925

527 Regarding the Arabic dataset, Table 4 shows that the best result was obtained by XGBoost (0.9259),
528 while the second-best result was yielded by RF (1.9750). In terms of *MAE*, XGBoost also achieved the
529 best result (0.4167), followed by NB (0.4876).
530 This outcome confirms that XGBoost performs better than the other classifiers; thus, it was used
531 in the next phase. Specifically, XGBoost outperformed the second-best algorithm in the English and
532 Arabic datasets with *RMSE* values of 1.4901 and 1.0491, respectively. XGBoost excels due to its effective
533 implementation of stochastic gradient boosting as well as its inbuilt regularization tools that prevent
534 overfitting.

Table 4. Results for the Arabic dataset.

Algorithm	*RMSE*	*MAE*
J48	2.2233	0.8396
RF	1.9750	0.7018
k-NN	2.1893	0.6534
NB	2.1611	0.4876
XGB	0.9259	0.4167

535 **Phase 2: Comparison with different metaheuristic algorithms combined with the XGB**
536 Several metaheuristic algorithms were compared after being combined with XGB in order to identify the
537 best combination. Three algorithms were used in this phase, namely, PSO, WOA, and MVO, each of
538 which has shown excellent results in the literature for different problems (*Ala'M et al.*, 2018; *Abd Elaziz*
539 *et al.*, 2019; *Rostami et al.*, 2020).
540 As shown in Table 5, PSO-XGB produced the best RMSE of 1.0993, while the MVO-XGB achieved
541 the best-second result (1.1168) and WOA-XGB yielded the worst result (1.1188). In terms of MAE,
542 PSO-XGB also obtained the best result (0.9258), followed by MVO-XGB (1.0520) and WOA-XGB
543 (1.0390).
544 Moreover, PSO-XGB showed better performance than the other algorithms in both measures, indicat-
545 ing that it is the best algorithm for this problem for the English dataset. The convergence of the three
546 algorithms can be seen in Fig. 2.

Table 5. Results for PSO, WOA, and MVO on the English dataset

Algorithm	*RMSE*	*MAE*
PSO-XGB	1.0993	0.9528
WOA-XGB	1.1188	1.0390
MVO-XGB	1.1168	1.0520
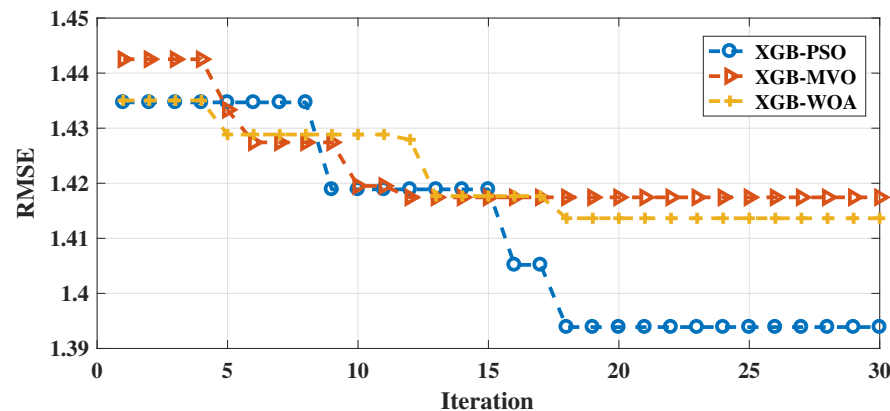
Figure 2. Convergence for PSO-XGB, MVO-XGB and WOA-XGB on the English dataset

547   Regarding the best values of the XGB parameters, the PSO-XGB (the superior model) selection can
548 be found in Table 6, where, for example, the min_child_weight is equal to 2 and γ is equal to 3.42. The
549 rest of the values are listed in the table.
550   As can be seen, the metaheuristic algorithms' selection of the best parameters enhanced the results in
551 all measures when compared with the first phase. This shows that the XGBoost parameters had a huge
552 impact on its performance. Therefore, this problem requires the use of metaheuristic algorithms.

Table 6. Best parameters for the English dataset

Parameters	Best value
min_child_weight	2
gamma γ	3.42
subsample	0.96
colsample_bytree	1
max_depth	2
learning_rate	0.91
n_estimators	10

553   As with the English dataset, three algorithms (PSO-XGB, WOA-XGB, and MVO-XGB) were com-
554 pared for the Arabic dataset. As shown in Table 7, the lowest *RMSE* result (0.7858) was obtained by the
555 PSO-XGB algorithm. The second-lowest value was achieved by the MVO-XGB algorithm (0.8896); this
556 is unlike the English dataset, for which this algorithm produced the worst result. For the *MAE* measure,
557 the best result was obtained also by the PSO-XGB algorithm (0.3999), followed by the GB-MVO and
558 WOA-XGB algorithms. The convergence of the three algorithms can be seen in Fig. 3.. As mentioned
559 earlier, the metaheuristic algorithms also improved the results for both measures since the best parameters
560 were selected for XGBoost.
561   The best values of the XGBoost parameters that were selected by PSO-XGB can be found in Table 8.

562 **Feature importance analysis**
563 Additional analyses were performed to identify the most important features (words). These keywords are
564 considered the top influencing features for each dataset in predicting the ratings of the reviews. Feature

Table 7. Results for PSO, WOA, and MVO on the Arabic dataset

Eng	*RMSE*	*MAE*
PSO-XGB	0.7722	0.3999
WOA-XGB	0.8531	0.4717
MVO-XGB	0.8186	0.4377
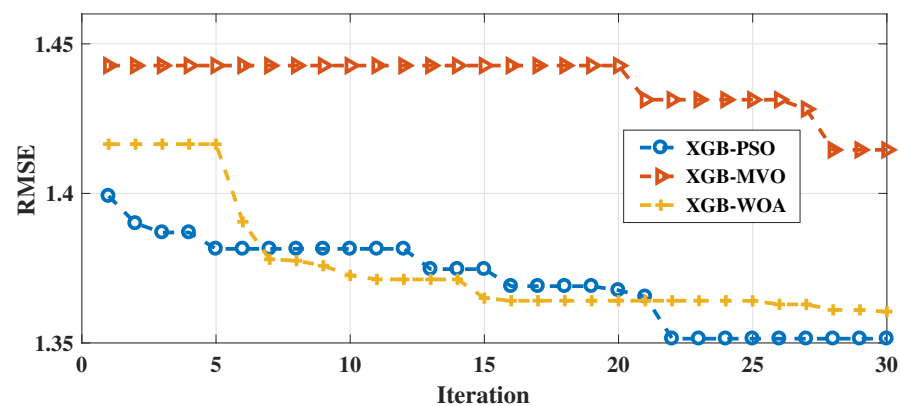
Figure 3. Convergence for PSO-XGB, MVO-XGB and WOA-XGB on the Arabic dataset

Table 8. Best parameters for the Arabic dataset

Parameters	Best value
min_child_weight	1
gamma γ	5
subsample	1
colsample_bytree	0.54
max_depth	12
learning_rate	0.72
n_estimators	510

importance was performed using the XGBoost algorithm. The mechanism employed for this task was calculated and weighted by the number of times the feature appeared in the tree structure (*Manju et al.*, 2019). The number of split points for each attribute responsible for improving the performance measure was used to determine the weights of features. These split points were defined by the Gini index method, and the average of all features' importance was calculated across all trees in the model.

Fig. 4 illustrates the top 20 features or words for both datasets. These features directly indicate consumers' opinions. Some words show similarity in importance, while others depict different views or thoughts.

The first five features for the Arabic dataset are F32, F69, F356, F124, and F224. Their translations are "taste," "prices," "employee," "a lot," and "great," respectively. The first feature (taste) indicates how important the sense of taste is to consumers and the extent to which the taste of food is considered an essential factor when choosing between restaurants. The second feature (prices) demonstrates the cost of the meal and if such a price is justified (high or low). The third feature (employee) indicates satisfaction with the service provided bystaff. The fourth feature (a lot) does not suggest any exact meaning other than the amount of something . The fifth feature (great) implies the adjective's positive meaning but without specifying which characteristics it refers to. Nevertheless, such a feature (great) is used by consumers when they emphasize their opinions, whether their review is positive or negative.

Meanwhile, the top five features for the English dataset are "price," "service," "bad," "great," and "late." The most important feature was the price, as the cost is crucial to consumers when they judge a meal or restaurant. Concerning the second feature, the service betokens how much can be significant to rate a restaurant. This indicates that excellent or horrible service is considered critical in the selection process. The third and fourth features both denote the quality of either the price, food, or service of a restaurant. Similar to the "great" feature, "bad" emphasizes the sentiment of the review. Regarding the fifth feature, receiving the food late is considered important, especially when a customer gives a low rating.

Moreover, both datasets show similarities and differences in the order of the features. Features such as "price," "taste," "late," "service," "cold," and "delivery" are similar, whereas "expectations," "overrated," and "spicy" show differences.

Top 20 (Arabic)			#	Top 20 (English)	
F#	**Term**	**Translation**	**-**	**F#**	**Term**
F32	طعم	taste	1	F99	price
F69	اسعار	prices	2	F102	service
F356	موظف	employee	3	F8	bad
F124	كثير	a lot	4	F243	great
F224	رائع	great	5	F160	late
F434	استلام	receive	6	F175	delicious
F98	شكرا	thanks	7	F310	best
F250	تاخير	late	8	F167	uneatable
F167	وصل	arrive	9	F148	hot
F260	زاكي	tasty	10	F218	Loved
F268	توصيل	delivery	11	F13	cold
F87	ساعة	hour (time)	12	F80	felt
F93	سيء	bad	13	F335	overrated
F213	حار	spicy	14	F221	amazing
F105	الخدمة	service	15	F28	order
F170	نص	half	16	F201	taste
F42	بارد	cold	17	F116	good
F657	ممتاز	excellent	18	F368	thanks
F629	شاورما	shawarma	19	F176	delivery
F191	طيب	delicious	20	F21	expectations

Figure 4. Top 20 features for English and Arabic datasets

Phase 3: Different classifiers combined with the best metaheuristic algorithms

In this final phase, PSO-XGB was compared with a well-known classifier (SVM) and combined with the PSO algorithm. As can be seen in Table 9, PSO-XGB achieved the best performance in terms of *RMSE*.

596 In this final phase, PSO-XGB was compared against a well-known classifier (SVM) in the literature
597 and combined with the PSO algorithm. As can be noticed in Table 9, PSO-XGB achieved the best
598 performance in terms of *RMSE*.

Table 9. Results for XGB and SVM combined with PSO for the English dataset

Algo	*RMSE*
PSO-XGB	1.0993
PSO-SVM	1.4204

599 The Arabic dataset results are shown in Table 10. The PSO-XGB algorithm outperformed the
600 PSO-SVM algorithm in terms of *RMSE*.

Table 10. Results for XGB and SVM combined with PSO for the Arabic dataset

Algo	*RMSE*
PSO-XGB	0.7722
PSO-SVM	0.9988

601 Furthermore, the 10-fold results in Table 11 for both datasets using PSO-XGB demonstrate significant
602 improvements. Additionally, s the statistical test (p-value) indicates that, in comparison with PSO-SVM,
603 the results for PSO-XGB are very small, suggesting strong statistical significance (Table 12). For instance,
604 if the p-value is 0.05 or less, it confirms that the observed differences are not due to random chance but
605 reflect a real improvement in performance.
606 Furthermore, the 10-fold results found in Table 11 for both datasets using PSO-XGB demonstrate
607 significant improvements, with the best results highlighted in italic font.

Table 11. The 10-fold results of the PSO-XGB for both English and Arabic datasets (Italic results are the best results).

10-Folds	English	Arabic
	PSO-XGB	
Fold 1	1.1100	0.7853
Fold 2	1.1155	0.7905
Fold 3	1.1309	0.8051
Fold 4	1.1256	0.8153
Fold 5	*1.0993*	0.8004
Fold 6	1.1401	0.7950
Fold 7	1.1452	0.7801
Fold 8	1.1208	*0.7722*
Fold 9	1.1354	0.8100
Fold 10	1.1500	0.8202
Average	**1.12728**	**0.79741**

Table 12. Statistical test (p-value) comparing PSO-XGB with PSO-SVM

Data	PSO-SVM
Enlgish	1.83E-04
Arabic	2.13E-34

Analysis of the error for PSO-XGB

609 Figures 5 and 6 show the errors of the predicted ordinal classes through the confusion matrix. These
610 errors explain the difference between each class and the class adjacent to it.
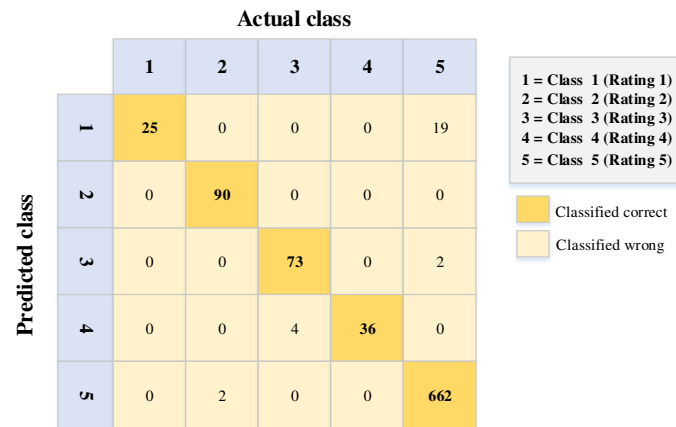
Figure 5. Confusion matrix values for the English dataset
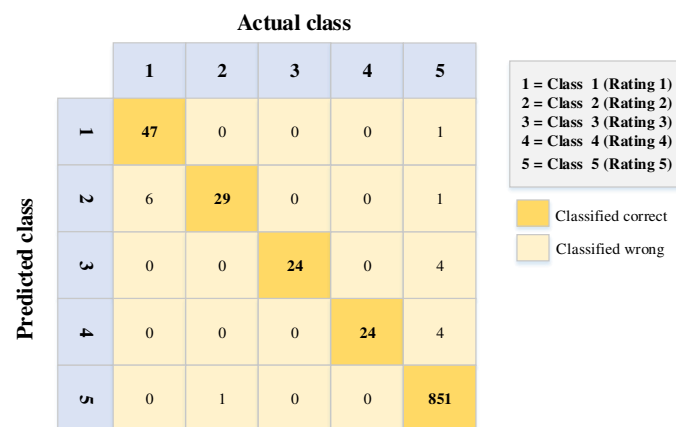
Figure 6. Confusion matrix values for the Arabic dataset

The data in Fig. Fig 6 (Arabic language) show better performance in terms of sentiment ordinal regression prediction than Fig. 5 (English language). In other words, for example, instances like rating (1) and (2) classified (error) as rating (5) in the English language, unlike the Arabic language, where there is less error in this matter. Conversely, 19 instances were classified as a rating of 1 when they should have been given a rating of 5. Again, this kind of error is less common in the Arabic data (Fig. 6).

After an extensive review of these errors (English data), it was determined that they occur for several reasons (ranked according to the majority), including:

- Irony and sarcasm. Some users use positive words to describe negative opinions. For example, one person had a rough day and expected to eat delicious food from a restaurant; however, he did not like the food. His review, which accompanied a (1) rating, stated, "That's just what I needed today!"

- Word ambiguity and choosing the wrong words to describe something (occurred when reviewers left reviews in a language other than their native language).

- Incorrect selection of the rating (occurred if the rater was confused about the meanings of ratings).

The findings indicate that it is hard to capture such reviews in the NLP model. However, this extended analysis explains the precise nature of such errors. Therefore, on such websites, restaurant owners should

627 be more careful when dealing with English reviews, as the reviewers are non-native English speakers.

628 Moreover, the accuracies for the confusion matrices of the PSO-XGBoost algorithm (Figs. 6 & 5) are

629 0.970 and 0.982 for the English and Arabic datasets, respectively. However, such results are not relevant

630 for this kind of problem (ordinal regression), since the weight of misclassification is not detected.

CONCLUSION

632 This work presented an ordinal regression sentiment polarity approach using the PSO-XGBoost algorithm

633 to assess restaurant reviews. Two types of pre-processing procedures were handled—one for each language

634 dataset (Arabic and English). Furthermore, the PSO algorithm functioned as an identifier and optimization

635 technique for the XGBoost parameters; it determined the optimal combination and eventually yielded

636 the best possible performance. It obtained superior results while handling complex tasks such as ordinal

637 regression problems (e.g., restaurant reviews).

638 The proposed approach was compared with other methods in three phases: first with standard classifiers

639 (J48, RF, k-NN, NB), then with other recent metaheuristic algorithms (MVO and WOA), and finally with

640 SVM. The proposed approach achieved better results than other methods in all phases. More specifically,

641 within the English dataset, PSO-XGB achieved an RMSE of 1.0993, outperforming WOA-XGB (1.1188),

642 MOV-XGB (1.1168), and PSO-SVM (1.420). Regarding the Arabic dataset, the proposed PSO-XGB

643 yielded an RMSE of 0.7722, meaning it outperformed WOA-XGB (0.8531), MOV-XGB (0.8186), and

644 PSO-SVM (0.9988).

645 Regarding the research questions, the proposed method can assist restaurant owners and provide early

646 alerts and feedback, allowing owners to focus on the most important terms (features) without having to

647 read all reviews in both languages. This, in turn, enables them to make better 672 business decisions

648 by utilizing relevant information. The method also reminds business owners to be cautious 673 when

649 handling non-native English speakers' reviewers. Further, the work achieved advanced performance 674

650 using the evolutionary XGBoost criteria, which performed better than state-of-the-art criteria.

651 Future research should implement more sophisticated model that can capture systematic reviews. 676

652 In addition, sampling more data would allow more terms and features to 677 be mapped with labels based

653 on the tree classification algorithm in XGBoost. Moreover, attention should 678 be paid to detecting

654 irony and sarcasm when assessing reviews. Doing so would ensure a comprehensive 679 understanding

655 of reviewers' true sentiments and help avoid misinterpretations that could impact decision-making and

656 customer satisfaction. As a result, the study's domain knowledge will improve, allowing it to be applied to

657 various fields, including predictive text and other approaches to natural language processing. Finally, more

658 comparisons with other algorithms can be conducted to compare the running times of various measures.

REFERENCES

660 Abd Elaziz, M., Oliva, D., Ewees, A. A., and Xiong, S. (2019). Multi-level thresholding-based grey scale

661 image segmentation using multi-objective multi-verse optimizer. *Expert Systems with Applications*,

662 125:112–129.

663 Adak, A., Pradhan, B., and Shukla, N. (2022). Sentiment analysis of customer reviews of food deliv-

664 ery services using deep learning and explainable artificial intelligence: Systematic review. *Foods*,

665 11(10):1500.

666 Adnan, M., Sarno, R., and Sungkono, K. R. (2019). Sentiment analysis of restaurant review with

667 classification approach in the decision tree-j48 algorithm. In *2019 International Seminar on Application*

668 *for Technology of Information and Communication (iSemantic)*, pages 121–126. IEEE.

669 Al Ameed, H., Al Ketbi, S., Al Kaabi, A., Al Shebli, K., Al Shamsi, N., Al Nuaimi, N., and Al Muhairi, S.

670 (2005). Arabic light stemmer: A new enhanced approach. In *The Second International Conference on*

671 *Innovations in Information Technology (IIT'05)*, pages 1–9.

672 Al-Qudah, D. A., Ala'M, A.-Z., Castillo-Valdivieso, P. A., and Faris, H. (2020). Sentiment analysis for

673 e-payment service providers using evolutionary extreme gradient boosting. *IEEE Access*, 8:189930–

674 189944.

675 Al-Zoubi, A., Alqatawna, J., Faris, H., and Hassonah, M. A. (2019). Spam profiles detection on social

676 networks using computational intelligence methods: The effect of the lingual context. *Journal of*

677 *Information Science*, page 0165551519861599.

Ala'M, A.-Z., Faris, H., Alqatawna, J., and Hassonah, M. A. (2018). Evolving support vector machines using whale optimization algorithm for spam profiles detection on online social networks in different lingual contexts. *Knowledge-Based Systems*, 153:91–104.

Ala'M, A.-Z., Heidari, A. A., Habib, M., Faris, H., Aljarah, I., and Hassonah, M. A. (2020). Salp chain-based optimization of support vector machines and feature weighting for medical diagnostic information systems. In *Evolutionary machine learning techniques*, pages 11–34. Springer.

AlZu'bi, S., Abu Zitar, R., Hawashin, B., Abu Shanab, S., Zraiqat, A., Mughaid, A., Almotairi, K. H., and Abualigah, L. (2022). A novel deep learning technique for detecting emotional impact in online education. *Electronics*, 11(18):2964.

Basiri, M. E., Ghasem-Aghaee, N., and Aghdam, M. H. (2008). Using ant colony optimization-based selected features for predicting post-synaptic activity in proteins. In *European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*, pages 12–23. Springer.

Bhoi, A. and Joshi, S. (2018). Various approaches to aspect-based sentiment analysis. *arXiv preprint arXiv:1805.01984*.

Bürkner, P.-C. and Vuorre, M. (2019). Ordinal regression models in psychology: A tutorial. *Advances in Methods and Practices in Psychological Science*, 2(1):77–101.

Cardoso, J. S., da Costa, J. F. P., and Cardoso, M. J. (2005). Modelling ordinal relations with svms: An application to objective aesthetic evaluation of breast cancer conservative treatment. *Neural Networks*, 18(5-6):808–817.

Chen, C., Zhang, Q., Yu, B., Yu, Z., Lawrence, P. J., Ma, Q., and Zhang, Y. (2020). Improving protein-protein interactions prediction accuracy using xgboost feature selection and stacked ensemble classifier. *Computers in Biology and Medicine*, 123:103899.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM.

Dashtipour, K., Poria, S., Hussain, A., Cambria, E., Hawalah, A. Y., Gelbukh, A., and Zhou, Q. (2016). Multilingual sentiment analysis: state of the art and independent comparison of techniques. *Cognitive computation*, 8(4):757–771.

Ding, Y., Zhou, K., and Bi, W. (2020). Feature selection based on hybridization of genetic algorithm and competitive swarm optimizer. *Soft Computing*, pages 1–10.

Dong, Z. and Dong, Q. (2006). *Hownet and the computation of meaning (with Cd-rom)*. World Scientific.

Faris, H., Alqatawna, J., Ala'M, A.-Z., and Aljarah, I. (2017). Improving email spam detection using content based feature engineering approach. In *2017 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)*, pages 1–6. IEEE.

Gaudette, L. and Japkowicz, N. (2009). Evaluation methods for ordinal classification. In *Canadian conference on artificial intelligence*, pages 207–210. Springer.

Gautam, G. and Yadav, D. (2014). Sentiment analysis of twitter data using machine learning approaches and semantic analysis. In *2014 Seventh International Conference on Contemporary Computing (IC3)*, pages 437–442. IEEE.

Ghamisi, P. and Benediktsson, J. A. (2014). Feature selection based on hybridization of genetic algorithm and particle swarm optimization. *IEEE Geoscience and remote sensing letters*, 12(2):309–313.

Habib, M., Faris, H., Hassonah, M. A., Alqatawna, J., Sheta, A. F., and Ala'M, A.-Z. (2018). Automatic email spam detection using genetic programming with smote. In *2018 Fifth HCT Information Technology Trends (ITT)*, pages 185–190. IEEE.

Hassonah, M. A., Al-Sayyed, R., Rodan, A., Ala'M, A.-Z., Aljarah, I., and Faris, H. (2020). An efficient hybrid filter and evolutionary wrapper approach for sentiment analysis of various topics on twitter. *Knowledge-Based Systems*, 192:105353.

Herbrich, R., Graepel, T., and Obermayer, K. (1999). *Regression models for ordinal data: A machine learning approach*. Citeseer.

Huang, F., Wei, K., Weng, J., and Li, Z. (2020). Attention-based modality-gated networks for image-text sentiment analysis. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(3):1–19.

Huang, F., Zhang, X., Zhao, Z., Xu, J., and Li, Z. (2019). Image–text sentiment analysis via deep multimodal attentive fusion. *Knowledge-Based Systems*, 167:26–37.

Jabreel, M. and Moreno, A. (2018). Eitaka at semeval-2018 task 1: An ensemble of n-channels convnet and xgboost regressors for emotion analysis of tweets. *arXiv preprint arXiv:1802.09233*.

Jiang, Y., Tong, G., Yin, H., and Xiong, N. (2019). A pedestrian detection method based on genetic algorithm for optimize xgboost training parameters. *IEEE Access*, 7:118310–118321.

Kapukaranov, B. and Nakov, P. (2015). Fine-grained sentiment analysis for movie reviews in bulgarian. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 266–274.

Karsi, R., Zaim, M., and El Alami, J. (2017). Impact of corpus domain for sentiment classification: An evaluation study using supervised machine learning techniques. In *Journal of Physics: Conference Series*, volume 870, page 12005.

Kauer, A. U. and Moreira, V. P. (2016). Using information retrieval for sentiment polarity prediction. *Expert Systems with Applications*, 61:282–289.

Kennedy, J. and Eberhart, R. (1995). Particle swarm optimization. In *Proceedings of ICNN'95-International Conference on Neural Networks*, volume 4, pages 1942–1948. IEEE.

Kern, B. M., Baumann, A., Kolb, T. E., Sekanina, K., Hofmann, K., Wissik, T., and Neidhardt, J. (2021). A review and cluster analysis of german polarity resources for sentiment analysis. In *3rd Conference on Language, Data and Knowledge (LDK 2021)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Krishna, A., Akhilesh, V., Aich, A., and Hegde, C. (2019). Sentiment analysis of restaurant reviews using machine learning techniques. In *Emerging Research in Electronics, Computer Science and Technology*, pages 687–696. Springer.

Le, L. T., Nguyen, H., Zhou, J., Dou, J., and Moayedi, H. (2019). Estimating the heating load of buildings for smart city planning using a novel artificial intelligence technique pso-xgboost. *Applied Sciences*, 9(13):2714.

Li, D., Wang, X., and Dey, D. (2019). Power link functions in an ordinal regression model with gaussian process priors. *Environmetrics*, 30(6):e2564.

Lin, S.-W., Ying, K.-C., Chen, S.-C., and Lee, Z.-J. (2008). Particle swarm optimization for parameter determination and feature selection of support vector machines. *Expert systems with applications*, 35(4):1817–1824.

Loke, R., Kachaniuk, O., Hammoudi, S., Quix, C., and Bernardino, J. (2020). Sentiment polarity classification of corporate review data with a bidirectional long-short term memory (bilstm) neural network architecture. In *DATA*, pages 310–317.

Manju, N., Harish, B., and Prajwal, V. (2019). Ensemble feature selection and classification of internet traffic using xgboost classifier. *International Journal of Computer Network & Information Security*, 11(7).

Martín-Valdivia, M.-T., Martínez-Cámara, E., Perea-Ortega, J.-M., and Ureña-López, L. A. (2013). Sentiment polarity detection in spanish reviews combining supervised and unsupervised approaches. *Expert Systems with Applications*, 40(10):3934–3942.

Mathieson, M. J. (1996). Ordinal models for neural networks. *Neural networks in financial engineering*, pages 523–536.

Mirjalili, S. and Lewis, A. (2016). The whale optimization algorithm. *Advances in engineering software*, 95:51–67.

Mirjalili, S., Mirjalili, S. M., and Hatamlou, A. (2016). Multi-verse optimizer: a nature-inspired algorithm for global optimization. *Neural Computing and Applications*, 27(2):495–513.

Naeem, M. Z., Rustam, F., Mehmood, A., Ashraf, I., and Choi, G. S. (2022). Classification of movie reviews using term frequency-inverse document frequency and optimized machine learning algorithms. *PeerJ Computer Science*, 8:e914.

Neethu, M. and Rajasree, R. (2013). Sentiment analysis in twitter using machine learning techniques. In *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*, pages 1–5. IEEE.

Nobre, J. and Neves, R. F. (2019). Combining principal component analysis, discrete wavelet transform and xgboost to trade in the financial markets. *Expert Systems with Applications*, 125:181–194.

Porter, M. F. (2001). Snowball: A language for stemming algorithms.

Rafique, A., Rustam, F., Narra, M., Mehmood, A., Lee, E., and Ashraf, I. (2022). Comparative analysis of machine learning methods to detect fake news in an urdu language corpus. *PeerJ Computer Science*, 8:e1004.

Ravi, K. and Ravi, V. (2015). A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-Based Systems*, 89:14–46.

Roh, M. and Park, K. (2019). Adoption of o2o food delivery services in south korea: The moderating role of moral obligation in meal preparation. *International Journal of Information Management*, 47:262 – 273.

Rostami, M., Forouzandeh, S., Berahmand, K., and Soltani, M. (2020). Integration of multi-objective pso based feature selection and node centrality for medical datasets. *Genomics*, 112(6):4370–4384.

Rupapara, V., Rustam, F., Amaar, A., Washington, P. B., Lee, E., and Ashraf, I. (2021). Deepfake tweets classification using stacked bi-lstm and words embedding. *PeerJ Computer Science*, 7:e745.

Saad, S. E. and Yang, J. (2019). Twitter sentiment analysis based on ordinal regression. *IEEE Access*, 7:163677–163685.

Shahi, T., Sitaula, C., and Paudel, N. (2022). A hybrid feature extraction method for nepali covid-19-related tweets classification. *Computational Intelligence and Neuroscience*, 2022.

Shao, C., Paynabar, K., Kim, T. H., Jin, J. J., Hu, S. J., Spicer, J. P., Wang, H., and Abell, J. A. (2013). Feature selection for manufacturing process monitoring using cross-validation. *Journal of Manufacturing Systems*, 32(4):550–555.

Shashua, A. and Levin, A. (2003). Ranking with large margin principle: Two approaches. In *Advances in neural information processing systems*, pages 961–968.

Shi, Y., Li, P., Yu, X., Wang, H., and Niu, L. (2018). Evaluating doctor performance: Ordinal regression-based approach. *Journal of medical Internet research*, 20(7):e240.

Song, K., Yan, F., Ding, T., Gao, L., and Lu, S. (2020). A steel property optimization model based on the xgboost algorithm and improved pso. *Computational Materials Science*, 174:109472.

Sun, L., Guo, J., and Zhu, Y. (2019). Applying uncertainty theory into the restaurant recommender system based on sentiment analysis of online chinese reviews. *World Wide Web*, 22(1):83–100.

Talabat (2004). Order food and grocery online from delivery restaurants and groceries in uae.

Thuseethan, S., Janarthan, S., Rajasegarar, S., Kumari, P., and Yearwood, J. (2020). Multimodal deep learning framework for sentiment analysis from text-image web data. In *2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, pages 267–274. IEEE.

Vinodhini, G. and Chandrasekaran, R. (2012). Sentiment analysis and opinion mining: a survey. *International Journal*, 2(6):282–292.

Wang, P., Tang, H., Zhang, H., Whiteaker, J., Paulovich, A. G., and Mcintosh, M. (2006). Normalization regarding non-random missing values in high-throughput mass spectrometry data. In *Biocomputing 2006*, pages 315–326. World Scientific.

Xia, J., Zhang, S., Cai, G., Li, L., Pan, Q., Yan, J., and Ning, G. (2017). Adjusted weight voting algorithm for random forests in handling missing values. *Pattern Recognition*, 69:52–60.

Yıldırım, P., Birant, U. K., and Birant, D. (2019). Eboc: Ensemble-based ordinal classification in transportation. *Journal of Advanced Transportation*, 2019.

Zhang, Y., Zhang, M., Liu, Y., and Ma, S. (2015). Boost phrase-level polarity labelling with review-level sentiment classification. *arXiv preprint arXiv:1502.03322*.