IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received October 2, 2019, accepted October 28, 2019, date of publication November 7, 2019, date of current version November 20, 2019.

*Digital Object Identifier 10.1109/ACCESS.2019.2952127*

# Twitter Sentiment Analysis Based on Ordinal Regression

**SHIHAB ELBAGIR[1,2] AND JING YANG[1]**
[1]College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China
[2]Faculty of Computer Science and Information Technology, Shendi University, Shendi 142-143, Sudan

Corresponding author: Shihab Elbagir (shihabsaad@yahoo.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61672179, Grant 61370083, and Grant 61402126, in part by the Heilongjiang Province Natural Science Foundation of China under Grant F2015030, in part by the Science Fund for Youths in Heilongjiang Province under Grant QC2016083, in part by the Postdoctoral Fellowship in Heilongjiang Province under Grant LBH-Z14071, and in part by the International Exchange Program of Harbin Engineering University for Innovation-Oriented Talents Cultivation.

**ABSTRACT** In recent years, research on Twitter sentiment analysis, which analyzes Twitter data (tweets) to extract user sentiments about a topic, has grown rapidly. Many researchers prefer the use of machine learning algorithms for such analysis. This study aims to perform a detailed sentiment analysis of tweets based on ordinal regression using machine learning techniques. The proposed approach consists of first pre-processing tweets and using a feature extraction method that creates an efficient feature. Then, under several classes, these features scoring and balancing. Multinomial logistic regression (SoftMax), Support Vector Regression (SVR), Decision Trees (DTs), and Random Forest (RF) algorithms are used for sentiment analysis classification in the proposed framework. For the actual implementation of this system, a twitter dataset publicly made available by the NLTK corpora resources is used. Experimental findings reveal that the proposed approach can detect ordinal regression using machine learning methods with good accuracy. Moreover, results indicate that Decision Trees obtains the best results outperforming all the other algorithms.

**INDEX TERMS** Machine learning technique, twitter, sentiment analysis, ordinal regression.

## I. INTRODUCTION

With the rapid development of social networks and microblogging websites. Microblogging websites have become one of the largest web destinations for people to express their thoughts, opinions, and attitudes about different topics [1], [2]. Twitter is a widely used microblogging platform and social networking service that generates a vast amount of information.

In recent years, researchers preferably made the use of social data for the sentiment analysis of people's opinions on a product, topic, or event. Sentiment analysis, also known as opinion mining, is an important natural language processing task. This process determines the sentiment orientation of a text as positive, negative, or neutral [3], [4].

Twitter sentiment analysis is currently a popular topic for research. Such analysis is useful because it gathers and classifies public opinion by analyzing big social data. However, Twitter data have certain characteristics that cause difficulty in conducting sentiment analysis in contrast to analyzing other types of data.

Tweets are restricted to 140 characters, written in informal English, contain irregular expressions, and contain several abbreviations and slang words. To address these problems, researchers have conducted studies focusing on sentiment analysis of tweets [5].

Twitter sentiment analysis approaches can be generally categorized into two main approaches, the machine learning approach, and a lexicon-based approach. In this study, we use machine learning techniques to tackle twitter sentiment analysis.

Most classification algorithms are focused on predicting nominal class data labels. However, a rule for predicting categories or labels on an ordinal scale involves many pattern recognition issues. This type of problem, known as ordinal classification or ordinal regression. Recently, ordinal regression has received considerable attention.

Ordinal regression issues in many fields of research are very common and have often been regarded as standard nominal problems that can lead to non-optimal solutions.

The associate editor coordinating the review of this manuscript and approving it for publication was Shuihua Wang.

In fact, Ordinal regression problems with some similarities and differences can be said to be between classification and regression. Medical research, age estimation, brain-computer interface, face recognition, facial beauty evaluation, image classification, social sciences, text classification, and more are some of the fields where ordinal regression is found.

Some studies suggest using machine learning techniques to solve regression problems to improve the sentiment analysis classification of Twitter data performance and predict new results. The main advantage of this method is the achievement of improved results.

The current study mainly focuses on the sentiment analysis of Twitter data (tweets) using different machine learning algorithms to deal with ordinal regression problems. In this paper, we propose an approach including pre-processing tweets, feature extraction methods, and constructing a scoring and balancing system, then using different techniques of machine learning to classify tweets under several classes.

The remainder of this paper is organized as follows: In section II, related studies about using different machine learning techniques for analyzing Twitter data are discussed. In section III, the methodology, data set, data preprocessing, features extracted, method of balancing and scoring and the machine learning techniques used are described. In section IV, the results and discussion of this study are presented. In section V, conclusions and recommendations for future work are provided.

## II. RELATED STUDIES

Various works have focused on analyzing social media data, especially those related to specific events. This intensive use of social media has attracted wide attention from academic research, and many investigations have been conducted to obtain important information on these events. In recent years, substantial studies have been carried out in the field of sentiment analysis on Twitter.

Jain and Dandannavar [6] examined several steps for sentiment analysis on Twitter data using machine learning algorithms. They also provided details of the proposed approach for sentiment analysis. The approach collected data and then preprocessed tweets using NLP-based techniques. Afterward, feature extraction was performed to export sentiment-relevant features. Finally, a model was trained using machine learning classifiers, such as naive Bayes classifiers, support vector machine (SVM), and decision tree. The proposed framework performed sentiment analysis using multinomial naive Bayes and decision tree algorithms. Results showed that decision trees perform effectively, showing 100% accuracy, precision, recall, and F1-score.

Substantial work has also been performed by Go *et al.* [7] who proposed a solution for sentiment analysis based on tweets using distant supervision. In their method, they used training data containing tweets with emoticons, which served as noisy labels. They built models using naive Bayes classifiers, maximum entropy (MaxEnt), and support vector machine. Their features comprised unigrams, bigrams, and POS. They concluded that SVM outperformed other models and that unigrams were more effective as features.

Along the same line, Bouazizi and Ohtsuki [8] introduced SENTA, which helps users select from a wide range of features those best suited to the application used to run the classification. The researchers used SENTA to analyze texts collected from Twitter's multi-class sentiment. The study was limited to seven different classes of sentiments. The results showed that the proposed approach reached an accuracy as high as 60.2% in the multi-classification. In both binary classification and ternary classification, this method has been shown to be sufficiently accurate.

Moreover, several SemEval works discussed the task of classifying the sentiment of tweets with hundreds of participants [9]–[11]. The evaluations are designed to investigate the essence of meaning in language as significance is intuitive for humans and thus it has proved elusive to transfer these intuitions to computational analysis.

There has been a growing interest in Sentiment Analysis based on Twitter data research as well as ordinal regression over the past decade. Ordinal regression problem is one of the main study areas in machine learning and data mining, with the aim of classifying patterns using a categorical scale showing a natural order between labels [12]–[14]. However, less attention was paid to the problems of ordinal regression (also known as ordinal classification). Recently, the field of ordinal regression has developed, many algorithms have been proposed from a machine learning approach for ordinal regression such as support vector ordinal regression and the perceptron ranking (PRank) algorithm.

There is a lot of research interest in studying ordinal regression problems. Gutiérrez *et al.* [15] highlighted the ordinal regression methods and proposes a taxonomy based on how models are designed to bring the order into consideration. A taxonomy of ordinal regression techniques divides them into three groups. In addition, a thorough experimental study is suggested to verify whether the use of order data improves the efficiency of the models, taking into account some of the taxonomy methods. The outcomes show that ordering information benefits ordinal models to enhance their accuracy and closeness of predictions to the ordinal scale objectives.

Li and Lin [16] proposed a reduction framework based on expanded examples from ordinal regression to binary classification. The framework can perform with any reasonable cost matrix and any binary classifier. The framework consists of three steps: removing expanded examples from the original examples, learning a binary classifier with any binary classification algorithm on the expanded examples, and building a binary classifier ranking rule. Their framework enables not only good ordinal regression algorithms based on well-tuned binary classification methods, but also new generalization boundaries for ordinal regression to be derived from recognized binary classification boundaries. Their framework also unifies many current ordinal regression algorithms.

Rennie and Srebro [17] introduced various loss functions in a systematic manner for discrete ordinal regression. Experiments carried out using the two threshold-based structures on the 1 Million MovieLens data set and discovered that the all-threshold building outperformed multi-class classification and simple regression techniques, as well as the immediate-threshold design. Two evident methods for managing discrete ordinal labels are: treating the distinct rating levels as unrelated classes and learning to forecast them as in a multi-class classification setting and treating them as real-valued answers and using a conventional regression setting with a loss function such as sum-squared error.

Niu *et al.* [18] present an End-to-End learning method to tackle ordinal regression problems using deep Convolutional Neural Network (CNN) that could simultaneously perform feature learning and modeling of regression. They apply it to the age estimation task and obtain better efficiency by avoiding designing hand-crafted characteristics directly. Estimating age can be cast as an ordinal regression problem. In particular, an ordinal regression problem is transformed into a series of binary classification sub-problems. And the authors propose a multiple output CNN learning algorithm to collectively solve these classification sub-problems so that the correlation between these tasks could be explored. They also release an Asian Face Age Dataset (AFAD) with more than 160 K facial images with accurate ground-age truths, which is the biggest dataset of public age to date.

In the recent few years, various work has been done in the field of Twitter sentiment analysis based on ordinal regression, but basically, all of those previous studies concentrated on transforming ordinal regression problems into a series of binary classification sub-problems. Whereas, to study the opinion of a user, it would be more interesting to go deeper into the classification and detecting a categorical scale of the sentiment hidden behind his post. Differences between our approach and the other researches discussed above are not small. However, note that our approach tackles sentiment analysis on tweets full of noise. Another advantage of our model is that our technique for sentiment polarity is also different as we deal with five categories of sentiment (highly positive, moderate positive, neutral, moderate negative, and highly negative) and therefore cannot apply their approach directly.

## III. METHODOLOGY

In this section, we present the methodology used in this study. The proposed system is basically composed of four main modules. The first module is data acquisition, which is a process of gathering labeled tweets to perform sentiment analysis; the second module, this dataset undergoes various steps of preprocessing to transform and refine tweets into a data set that can be easily used for subsequent analysis. The third module concerns the extraction of relevant features for building a classification model. Then, the balancing and scoring tweets technique is illustrated. The last module is applying different machine learning classifiers that classify the tweets into high positive, moderate positive, neutral, moderate negative, and high negative. Figure 1 shows the various steps performed for sentiment analysis using machine learning algorithms.

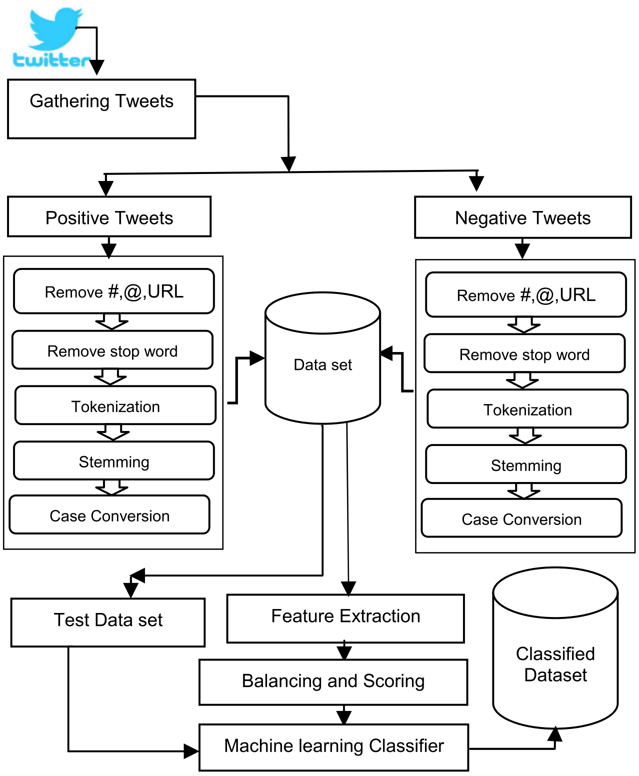


**FIGURE 1.** The system architecture.

**TABLE 1.** Data set summary.

| Dataset | Positive | Negative | Total |
|---|---|---|---|
| Training | 3500 | 3500 | 7000 |
| Test | 1500 | 1500 | 3000 |

### A. DATASET COLLECTION AND TWEETS PREPROCESSING

#### 1) DATASET COLLECTION

A dataset is collected from Twitter by using Twitter API, and the data are annotated as positive or negative tweets. The dataset is publicly available in the Natural Language Toolkit (NLTK) corpora resource, which is well-known and extensively analyzed in many studies. The total corpus size is 10000 twitter posts, consisting of 5000 positive tweets and 5000 negative tweets. For the sake of this task, we gathered and prepared data sets as follows: set contains 7000 tweets, this set is used for the training model. There are 3000 tweets in the other set, this set will serve as a test set. Table 1 shows a summary of the used datasets.

#### 2) PREPROCESSING OF TWEETS

Raw tweets are full of noise, misspellings, and contain numerous abbreviations and slang words. Such noisy characteristics often involve the performance of sentiment analysis approaches. Thus, some preprocessing approaches are applied prior to feature extraction. The preprocessing of tweets include the following steps:

- removing all hyperlinks ("http://url"), hashtags ("# hashtag"), retweet (RT), and username links ("@username") that appear in the tweets;
- removing repeated letters (e.g., "loovee" becomes "love");
- avoiding misspellings and slang words;
- converting words into lowercase (case conversion);
- removing commonly used words that do not have special meaning, such as pronouns, prepositions, and conjunctions (stop word removal);
- reducing words to their stem or common root by removing plurals, genders, and conjugation (stemming);
- segmenting sentences into words or phrases called token by discarding some characters (tokenization);
- removing duplicated data tweets; and
- replacing all emoticons with their corresponding sentiment.

Algorithm 1 displays the overall procedure used in the preprocessing of the data set in this study.

**Algorithm 1** Preprocessing Tweets

```
Begin
    Input Query String
    Until the data is retrieved from Twitter Streaming
    API, Do:
    Filter English Language Tweets
    Remove duplicate tweets
    Case Conversion
    For each tweet, Do:
        Procedure Pre-processing(tweet):
        Remove Twitter symbols(#topic, @user name,
        retweet (RT))
          Remove URL ("http://url")
          Remove all symbols, numbers, Emoticons,
          and punctuations
          Avoiding misspellings and slang words
          Remove repeated letters
          Remove Stop Words
          Tokenization
          Stemming
          Return tweet
    End Procedure
    Procedure Feature Extraction (clean_tweet):
        Extract Features using (TF-IDF) in a format
    suitable to machine learning algorithms
    End Procedure
    Procedure Balancing and scoring(Features):
        Calculating polarity degree of the tweets,‘
    balancing and labeling the tweets
    End Procedure
    Procedure Sentiment Classification (Feature):
        Classify tweet using machine learning
    techniques (SoftMax, SVR, RF, and DT)
    End Procedure
End Until
End
```

### B. FEATURE EXTRACTION

In the feature extraction phase, we extract aspects or features for building a classification model. The extracted features are in a format suitable to sustain directly to machine learning algorithms from datasets containing raw data of different formats, such as text, a sequence of symbols, and images. Therefore, we use certain techniques to extract features from the tweets, such as count vectorizer and term frequency-inverse document frequency (TF–IDF). In this study, we use TF-IDF to extract the feature matrix that reflects the importance of terms to the corpus in a text.

#### 1) TERM FREQUENCY–INVERSE DOCUMENT FREQUENCY (TF–IDF)

TF–IDF is a statistical measure and a term-weighting scheme that provides the bag-of-words model with information on word importance. Different aspects are extracted from the processed dataset, such as verbs, adjectives, and nouns. Subsequently, these aspects are used to calculate the sentiment polarity in a sentence to determine the opinion of individuals by using models, such as unigrams, bigrams, or n-grams [19].

TF–IDF is used to evaluate the significance of a word to a document in a dataset. Each word is assigned a weight in the document [20]. We use for that TF–IDF vectorizer Python module of Scikit-learn. A TF–IDF vectorizer extracts features based on word count, providing less weight to frequent words and more weight to rare words [21]. Table 2 presents the parameters used for feature extraction.

**TABLE 2.** TF-IDF parameter.

| Parameter | Value |
|---|---|
| Min.document frequency | 5 |
| Max.document frequency | 0.75 |
| N-gram range | 1,3 |
| Max features | 10000 |

The TF–IDF weight of a term within a document is Refer to "(1)", where $t$ is a term appearing in document $d$:

$$TF-IDF(t,d) = TF(t,d).IDF(t)$$
$$IDF(t) = \log\frac{N}{DF(t)} \tag{1}$$

where TF (*t, d*) is the number of times the term $t$ appears in a particular document $d$, and IDF is the total size of document N divided by the number of documents in the entire dataset D, which contains the term $t$.

### C. BALANCING AND SCORING METHOD

In order to build a sentiment analysis classification model, that has the main aim to analyze the sentiments of tweets and classify them as high positive, moderate positive, neutral, moderate negative, and high negative according to the polarity of the tweet. At the first stage of this work, after tokens are transformed into aspects or features using TF–IDF vectorizer. Then a new method for scoring and labeling tweets is then proposed. The method is consists of the following steps:

- Step 1: We assume that each tweet can be classified as positive and/or negative tweets depending on the number of positive or negative words in the tweet and based on the emoticons in the tweet. This way the polarity score of each tweet can be calculated. During the collection time, we ignored neutral tweets because it is difficult and could affect the results.
- Step 2: In this step, we determine the tweet polarity by summing up the value assigned to each feature stated in the tweet, for each tweet($t$) the overall polarity score value, is calculated as:

$$Polarity\ \ score\,(tweet) = \sum_{n=1}^{n} feature\ value\,(tweet) \tag{2}$$

  where: $n$ is the length of a tweet ($t$) (i.e., number of features in a single tweet ($t$)).
- Step 3: we define a rule for determining the overall sentiment polarity in a tweet in five sentiment classes. The polarity is categorized as follows: If $0 <$ polarity ($t$) $< 2$, then, the tweet is classified as moderate positive. If $2 \leq$ polarity ($t$) $\leq 4$, it is highly positive. If $-2 <$ polarity ($t$) $< 0$, then, the tweet is classified as moderate negative. If $-4 \leq$ polarity ($t$) $\leq -2$, it is highly negative. In the remaining case, if the polarity ($t$) is equal to 0, the tweet is classified as neutral.

$$\text{Tweet sentiment} = \begin{cases} \text{Moderate Positive if polarity(t)>0.0 and} \\ \text{polarity (t)} \ < 2.0 \\ \text{High Positive if polarity(t)} \ > 0 \text{ and} \\ \text{polarity (t)} \ <= 4.0 \\ \text{Moderate Negative if polarity(t)>-2 and} \\ \text{polarity(t)} \ < 0.0 \\ \text{High Negative if polarity(t)} \ >= -4 \text{ and} \\ \text{polarity (t)} \ <= -2 \\ \text{Neutral if polarity(t)} \ = 0.0 \end{cases} \tag{3}$$

- Step 4: For each class, we assign a score as follows: $+2$, $+1$, $0$, $-1$, and $-2$ respectively for high positive, moderate positive, neutral, moderate negative, and high negative.

$$\text{Score (t)} = \begin{cases} +2 \text{ if (Tweet sentiment) is classified as High} \\ \text{Positive} \\ +1 \text{ if (Tweet sentiment) is classified as} \\ \text{Moderate Positive} \\ -1 \text{ if (Tweet sentiment) is classified as} \\ \text{Moderate Negative} \\ -2 \text{ if (Tweet sentiment) is classified as High} \\ \text{Negative} \\ 0 \quad \text{if (Tweet sentiment) value is classified as} \\ \text{Neutral} \end{cases} \tag{4}$$

- Step 5: is the final step, in which we applying machine learning techniques in order to classify the overall sentiment of tweets.

## IV. MACHINE LEARNING TECHNIQUES

Two basic approaches for detecting sentiment analysis from the text (tweet), namely, Lexicon-based and machine learning approaches, are available [22]–[24]. In this study, we use a machine learning approach. Different types of machine learning techniques are used for text classification and twitter sentiment analysis. Machine learning techniques train the algorithm with some specific training data with known outputs, thereby allowing working with new test data. Several machine learning algorithms include multinomial logistic regression (SoftMax), Support Vector Regression (SVR), Decision Trees (DTs), and Random Forest (RF) used to build the study machine learning classifier.

### A. SUPPORT VECTOR REGRESSION

Support vector machine a set of supervised learning methods supports detection of classification, regression, and outliers that are helpful for statistical theory of learning. It is possible to extend support vector classification to solve regression problems. This technique is called Support Vector Regression (SVR). The SVR maintains all the main features of SVM and uses for classification the same principles as the SVM, with only a few minor changes. Support vector regression has three different implementations: SVR, Nu-SVR and Linear SVR [25], [26].

### B. DECISION TREE

Decision Trees (DTs) are a non-parametrically supervised learning algorithm that is commonly used for task classification and task regression. It works for categorical as well as continuous variables dependent. The aim is to create a model that predicts the value of a target variable by learning from the data inferred simple rules of decision. The training data space is displayed in a hierarchical representation in which the data is partitioned by a condition on the attribute value [27].

### C. RANDOM FOREST (RF)

Random Forest (RF) is a technique of classification and regression based on the ensemble method, which is based on the bagging of bootstraps [28]. Boosting and bagging are the two commonly known and used techniques for the classification of trees. The random forest consists of a combination of trees that can be used to predict the class label based on the categorical dependent variable for a specified data point [29]. Using a random subset of the original features, each decision tree is trained. To determine the sample class, a new sample is used to classify the mode of the outputs of each tree within the forest. This algorithm is used in many applications of speech and language processing.

### D. MULTINOMIAL LOGISTIC REGRESSION (SOFTMAX)

The extension of logistic regression for multiple classes is known as (SoftMax regression) [30]. SoftMax is highly similar to the MaxEnt model. However, the problem is explained in a different formula, and another algorithm in which each word has a different weight for each class is used to estimate the weights. We can represent the documents as sparse vectors with term frequencies or binary occurrence values. The SoftMax function is used in SoftMax regression, the input to the function is the outcome of distinct linear functions of $K$ and the estimated probability of a sample vector $x$ and a weighting vector $w$ for the $j$'th class is referred to "(5)", as follows:

$$P(y=j|x)\frac{e^{(x^T w_j)}}{\sum_{k=1}^{K} e^{(x^T w_k)}} \tag{5}$$

## V. RESULTS AND DISCUSSION

In the current study, we use four machine learning techniques to perform sentiment analysis of Twitter data based on ordinal regression. SoftMax, SVR, RF, and DT are the algorithms used for categorization. Experiments are conducted using Scikit-learn [31], [21], an open-source of machine learning software packages in Python. The Scikit-library provides various machine learning models to implement codes easily. The Twitter dataset is made publicly available by the NLTK corpora resources, familiar, and extensively analyzed in previous studies. After tweets preprocessing, applying feature extraction techniques, and scoring and balancing data, we run different tests using machine learning algorithms. The results of the preprocessing and feature extraction processes are displayed in Table 3.

Table 4 summarizes the distribution of the total number of tweets and the percentages for the five sentiment classes. Based on the outcomes shown in Table 4, it can be revealed that about (60%) of the tweets were moderate positive and moderate negative in sentiment, with 16.98% expressing high positive sentiment, 6.27% expressing neutral sentiment, and 15.38% expressing high negative sentiment. Moreover, the sentiment analysis polarity according to the count of tweets in each class is represented graphically in Figure 2.

**TABLE 3. Preprocessing tweets.**

| | Text | Label |
|---|---|---|
| 0 | Follow friday inte paris top engaged members co... | High Positive |
| 1 | listen last night bleed amazing track scotland | High Positive |
| 2 | congrats | moderate Positive |
| 3 | everyone gon na talking abt rat boy today bc | High Negative |
| 4 | earth assume rain london likely influenced ove... | High Negative |
| 5 | remember fab four 24 hour call damn miss much | High Negative |
| 6 | thirsty | Neutral |

**TABLE 4. Polarity of tweets.**

| Polarity | Total Tweets |
|---|---|
| High Positive | 1698 (16.98%) |
| Moderate Positive | 3043 (30.43%) |
| Neutral | 627 (6.27%) |
| Moderate Negative | 3094 (30.94%) |
| High Negative | 1538 (15.38%) |

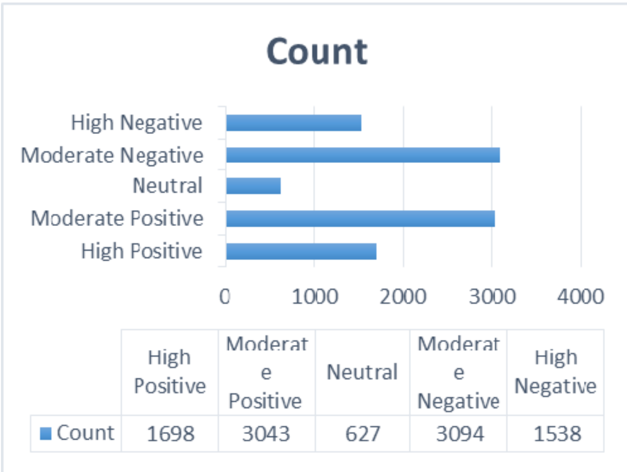


**FIGURE 2. Sentiment of tweets.**

Regarding the evaluation of the methods, the overall accuracy obtained by the 10-fold cross-validation in each of the classifiers is used as one of the main classification performance metrics. Table 5 shows the classification performance of all machine learning algorithms using the 10-fold cross-validation scores, which are used to partition data randomly into 10 subsets in which the class is represented in approximately the same proportions as in the full dataset.

Table 6 shows the results of the random forest algorithm classification. The accuracy obtained reaches 83.2%. Noticeably, the precision of high positive tweets is the highest at 95%. This implies that tweets categorized as high positive tend to be more correctly categorized. However, the recall of

**TABLE 5. 10-fold cross validation scores.**

| Algorithm | Score1 | Score2 | Score3 | Score4 | Score5 | Score6 | Score7 | Score8 | Score9 | Score10 |
|---|---|---|---|---|---|---|---|---|---|---|
| SoftMax | 0.660 | 0.625 | 0.674 | 0.627 | 0.657 | 0.648 | 0.682 | 0.637 | 0.687 | 0.664 |
| SVR | 0.828 | 0.819 | 0.812 | 0.819 | 0.812 | 0.820 | 0.812 | 0.791 | 0.828 | 0.827 |
| RF | 0.699 | 0.728 | 0.730 | 0.744 | 0.720 | 0.745 | 0.729 | 0.734 | 0.755 | 0.751 |
| DT | 0.863 | 0.868 | 0.838 | 0.858 | 0.853 | 0.871 | 0.869 | 0.860 | 0.864 | 0.882 |

**TABLE 6. Random Forest validation metrics.**

| | Precision | Recall | F1-score | support |
|---|---|---|---|---|
| High Negative | 0.92 | 0.69 | 0.79 | 455 |
| Moderate Negative | 0.86 | 0.89 | 0.88 | 937 |
| Neutral | 0.55 | 1.00 | 0.71 | 205 |
| Moderate Positive | 0.84 | 0.88 | 0.86 | 922 |
| High Positive | 0.95 | 0.69 | 0.80 | 481 |
| Avg / total | 0.86 | 0.83 | 0.83 | 3000 |

moderate negative tweets is the highest at 89%. The precision of neutral tweets is the lowest at 55%. This can be clarified by the reality that, after pre-processing step and removing hashtags and/or emoticons, many tweets from moderate negative or moderate positive classes might be classified as neutral.

The results of the SoftMax algorithm are displayed in Table 7. Despite the number of classes, accuracy reached 67.2%. However, precision reaches 62%. Noticeably, some sentiment classes seem to be easier to detect than others. Recall tweets belonging to the moderate positive and moderate negative class is the highest, reaching 92%. This shows that it is easy to distinguish tweets belonging to this class from other classes. A significant unbalance between the different classes can be noted. We also observe that the amount of training data is smaller for the neutral class than that for the other classes.

The outcome in Table 8 shows that the overall recall of Decision Trees is reached 85%. However, the precision of moderate negative tweets is the highest reaching 92%. This means that tweets which are classified as moderate negative are mostly negative. The accuracy obtained is equal to 91.81%.

Table 9 shows the results of the different algorithms evaluation metrics using F1-score. We observe that decision Tree performs better than Random Forest and multinomial logistic regression. It can also be observed that all three algorithms perform better for moderate positive and moderate negative tweets than they do for neutral. Moreover, Random Forest and Decision Tree perform better for high positive

**TABLE 7. Softmax validation metrics.**

| | Precision | Recall | F1-score | support |
|---|---|---|---|---|
| High Negative | 0.60 | 0.24 | 0.34 | 455 |
| Moderate Negative | 0.65 | 0.92 | 0.76 | 937 |
| Neutral | 0.00 | 0.00 | 0.00 | 205 |
| Moderate Positive | 0.70 | 0.92 | 0.79 | 922 |
| High Positive | 0.73 | 0.40 | 0.52 | 481 |
| Avg / total | 0.62 | 0.67 | 0.62 | 3000 |

**TABLE 8. Decision Trees classifier validation metrics.**

| | Precision | Recall | F1-score | support |
|---|---|---|---|---|
| High Negative | 0.86 | 0.83 | 0.85 | 465 |
| Moderate Negative | 0.92 | 0.85 | 0.88 | 939 |
| Neutral | 0.52 | 1.00 | 0.68 | 176 |
| Moderate Positive | 0.90 | 0.84 | 0.87 | 928 |
| High Positive | 0.86 | 0.82 | 0.84 | 492 |
| Avg / total | 0.87 | 0.85 | 0.85 | 3000 |

**TABLE 9. F1-score for classifiers.**

| Algorithm | High Negative | Moderate Negative | Neutral | Moderate Positive | High Positive | Total |
|---|---|---|---|---|---|---|
| Multinomial logistic regression | 0.34 | 0.76 | 0.00 | 0.79 | 0.52 | 0.62 |
| Random Forest | 0.79 | 0.88 | 0.71 | 0.86 | 0.80 | 0.83 |
| Decision Tree | 0.85 | 0.88 | 0.68 | 0.87 | 0.84 | 0.85 |

and high negative tweets than Multinomial logistic regression classifier.

Table 10 displays the overall prediction accuracy of all algorithms. Decision Tree and Random Forest achieve much better accuracy when compared to Support vector regression

**TABLE 10.** Accuracy of all classifiers.

| Algorithm | Accuracy |
|---|---|
| Multinomial logistic regression | 67.2% |
| Support Vector Regression | 81.95% |
| Random Forest | 83.2% |
| Decision Tree | 91.81% |

**TABLE 11.** Mean squared error and mean absolute error results.

| Algorithm | Mean Absolute Error (MAE) | Mean Squared Error (MSE) |
|---|---|---|
| Multinomial logistic regression | 0.342 | 0.328 |
| Support Vector Regression | 0.410 | 0.337 |
| Random Forest | 0.163 | 0.172 |
| Decision Tree | 0.154 | 0.155 |

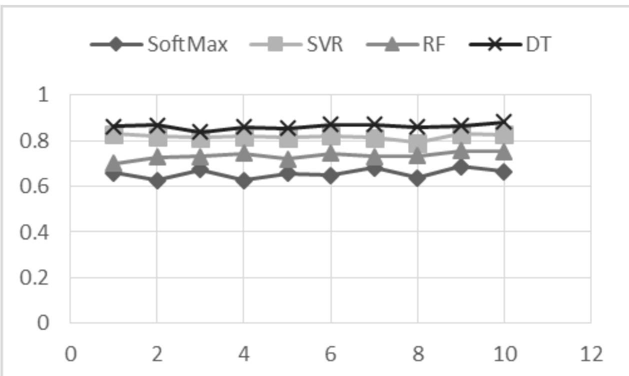


**FIGURE 3.** The scores of the 10-fold cross-validations.

and Multinomial logistic regression. Decision Tree gives the highest accuracy at 91.81%.

In order to evaluate the efficiency of the algorithms used for ordinal regression, several evaluation metrics are used. However, Mean Square Error (MSE) and Mean Absolute Error (MAE) are the most common ones. In contrast to the accuracy measure, lower values in MAE and MSE are better because MAE and MSE are an error measurement. The results of applying the Mean Squared Error (MSE) and Mean Absolute Error (MAE) are shown in Table 11.

The evaluation based on the Multinomial logistic regression (SoftMax), Support Vector Regression (SVR), Decision Tree (DT), and Random Forest (RF) algorithms using a 10-fold cross-validation method for testing is represented in Figure 3. Decision Tree produced the best score results and Support vector regression scores result exceeded multinomial logistic regression results by a considerable margin.

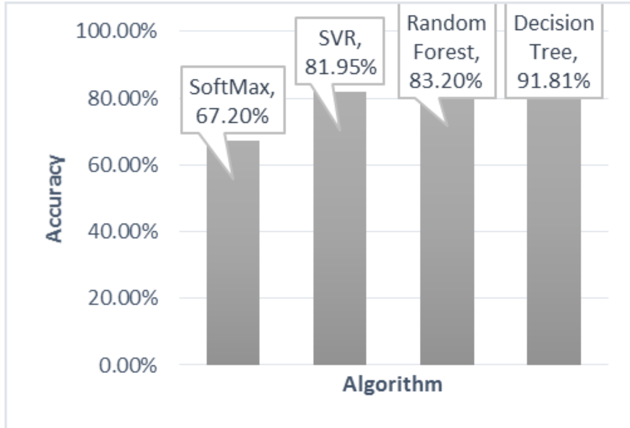


**FIGURE 4.** The accuracy of each algorithm.

Figure 4 graphically represents the prediction accuracy of SoftMax, SVR, DT, and RF techniques. It is observed that SoftMax has lower accuracy (67.2%), SVR and RF classifiers have achieved a good accuracy (81.95, 83.20% respectively), whereas DT has the highest accuracy among all other classifiers (91.81%).

## VI. CONCLUSION

This study aims to explain sentiment analysis of twitter data regarding ordinal regression using several machine learning techniques. In the context of this work, we present an approach that aims to extract Twitter sentiment analysis by building a balancing and scoring model, afterward, classifying tweets into several ordinal classes using machine learning classifiers. Classifiers, such as Multinomial logistic regression, Support vector regression, Decision Trees, and Random Forest, are used in this study. This approach is optimized using Twitter data set that is publicly available in the NLTK corpora resources.

Experimental results indicate that Support Vector Regression and Random Forest have an almost similar accuracy, which is better than that of the Multinomial logistic regression classifier. However, the Decision Tree gives the highest accuracy at 91.81%. Experimental results concluded that the proposed model can detect ordinal regression in Twitter using machine learning methods with a good accuracy result. The performance of the model is measured using accuracy, Mean Absolute Error, and Mean Squared Error.

In the future, we plan to improve our approach by attempting to use bigrams and trigrams. Furthermore, we intend to investigate different machine learning techniques and deep learning techniques, such as Deep Neural Networks, Convolutional Neural Networks, and Recurrent Neural Networks.

## REFERENCES


[1] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, "From tweets to polls: Linking text sentiment to public opinion time series.," in *Proc. ICWSM*, 2010, vol. 11, nos. 122–129, pp. 1–2.

[2] M. A. Cabanlit and K. J. Espinosa, "Optimizing N-gram based text feature selection in sentiment analysis for commercial products in Twitter through polarity lexicons," in *Proc. 5th Int. Conf. Inf., Intell., Syst. Appl. (IISA)*, Jul. 2014, pp. 94–97.

[3] S.-M. Kim and E. Hovy, "Determining the sentiment of opinions," in *Proc. 20th Int. Conf. Comput. Linguistics*, Aug. 2004, p. 1367.

[4] C. Whitelaw, N. Garg, and S. Argamon, "Using appraisal groups for sentiment analysis," in *Proc. 14th ACM Int. Conf. Inf. Knowl. Manage.*, Oct./Nov. 2005, pp. 625–631.

[5] H. Saif, M. Fernández, Y. He, and H. Alani, "Evaluation datasets for Twitter sentiment analysis: A survey and a new dataset, the STS-Gold," in *Proc. 1st Interantional Workshop Emotion Sentiment Social Expressive Media, Approaches Perspect. AI (ESSEM)*, Turin, Italy, Dec. 2013.

[6] A. P. Jain and P. Dandannavar, "Application of machine learning techniques to sentiment analysis," in *Proc. 2nd Int. Conf. Appl. Theor. Comput. Commun. Technol. (iCATccT)*, Jul. 2016, pp. 628–632.

[7] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," *Processing*, vol. 150, no. 12, pp. 1–6, 2009.

[8] M. Bouazizi and T. Ohtsuki, "A pattern-based approach for multi-class sentiment analysis in Twitter," *IEEE Access*, vol. 5, pp. 20617–20639, 2017.

[9] R. Sara, R. Alan, N. Preslav, and S. Veselin, "SemEval-2016 task 4: Sentiment analysis in Twitter," in *Proc. 8th Int. Workshop Semantic Eval.*, 2014, pp. 1–18.

[10] S. Rosenthal, P. Nakov, S. Kiritchenko, S. Mohammad, A. Ritter, and V. Stoyanov, "Semeval-2015 task 10: Sentiment analysis in Twitter," in *Proc. 9th Int. Workshop Semantic Eval. (SemEval)*, Jun. 2015, pp. 451–463.

[11] P. Nakov, A. Ritter, S. Rosenthal, F. Sebastiani, and V. Stoyanov, "SemEval-2016 task 4: Sentiment analysis in Twitter," in *Proc. 10th Int. Work. Semant. Eval.*, Jun. 2016, pp. 1–18.

[12] A. K. Jain, R. P. W. Duin, and J. C. Mao, "Statistical pattern recognition: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 1, pp. 4–37, Jan. 2000.

[13] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*. San Mateo, CA, USA: Morgan Kaufmann, 2016.

[14] V. Cherkassky and F. M. Mulier, *Learning From Data: Concepts, Theory, and Methods*. Hoboken, NJ, USA: Wiley, 2007.

[15] P. A. Gutiérrez, M. Pérez-Ortiz, J. Sánchez-Monedero, F. Fernández-Navarro, and C. Hervás-Martínez, "Ordinal regression methods: Survey and experimental study," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 1, pp. 127–146, Jan. 2016.

[16] L. Li and H. Lin, "Ordinal regression by extended binary classification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 865–872.

[17] J. D. Rennie and N. Srebro, "Loss functions for preference levels: Regression with discrete ordered labels," *IJCAI Work. Adv. Preference Handling*, Jul. 2005, pp. 180–186.

[18] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua, "Ordinal regression with multiple output CNN for age estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 4920–4928.

[19] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Oct. 2013, pp. 1631–1642.

[20] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval," *J. Documentation*, vol. 28, no. 1, pp. 11–21, 1972.

[21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.

[22] X. Chen, M. Vorvoreanu, and K. Madhavan, "Mining social media data for understanding students' learning experiences," *IEEE Trans. Learn. Technol.*, vol. 7, no. 3, pp. 246–259, Jul./Sep. 2014.

[23] N. R. Kasture and P. B. Bhilare, "An approach for sentiment analysis on social networking sites," in *Proc. Int. Conf. Comput. Commun. Control Autom.*, Feb. 2015, pp. 390–395.

[24] V. Singh and S. K. Dubey, "Opinion mining and analysis: A literature review," in *Proc. 5th Int. Conf.-Confluence Next Gener. Inf. Technol. Summit (Confluence)*, Sep. 2014, pp. 232–239.

[25] W. Chu and S. S. Keerthi, "Support vector ordinal regression," *Neural Comput.*, vol. 19, no. 3, pp. 792–815, 2007.

[26] S. Liu, F. Li, F. Li, X. Cheng, and H. Shen, "Adaptive co-training SVM for sentiment classification on tweets," in *Proc. 22nd ACM Int. Conf. Inf. Knowl. Manage.*, Oct./Nov. 2013, pp. 2079–2088.

[27] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986.

[28] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[29] A. Celikyilmaz, D. Hakkani-Tür, and J. Feng, "Probabilistic model-based sentiment analysis of Twitter messages," in *Proc. IEEE Spoken Lang. Technol. Workshop*, Dec. 2010, pp. 79–84.

[30] A. Genkin, D. D. Lewis, and D. Madigan, "Large-scale Bayesian logistic regression for text categorization," *Technometrics*, vol. 49, no. 3, pp. 291–304, 2007.

[31] *Scikit-Learn: Machine Learning in Python*. Accessed: Aug. 10, 2019. [Online]. Available: http://scikit-learn.org/stable/



**SHIHAB ELBAGIR** received the B.S. degree in computer science from Alneelain University, Sudan, in 2002, and the M.S. degree in computer science and information from the University of Gezira, Sudan, in 2007. He is currently pursuing the Ph.D. degree with the College of Computer Science and Technology, Harbin Engineering University, Harbin, China. He is currently a Lecturer with the Faculty of Computer Science and Information Technology, Shendi University, Shendi, Sudan.



**JING YANG** received the bachelor's degree in computer and application from Northeast Heavy Machinery College, in July 1984, and the Ph.D. degree in computer application technology from Harbin Engineering University, Harbin, China, in June 2006. She is currently a Professor and a Ph.D. Supervisor with the College of Computer Science and Technology, Harbin Engineering University, and the Deputy Director with the School of Computer Science and Technology, Institute of Software and Data Science. Since 1984, she has been engaged in the teaching and scientific research of computer science. More than 130 articles have been published in important academic journals and international conferences of IEEE at home and abroad, of which more than 100 articles have been retrieved by SCI, EI, and ISTP. Her current research interests include big data analysis and privacy protection, social networks, information security, and scheduling algorithm. She is also a Senior Member of CCF, member of the Communication of Chinese Information Processing Society, Harbin Information Expert, and the Head of the Data Structure Course of Heilongjiang Provincial Excellent Course.