



e-ISSN: 2319-8753 | p-ISSN: 2347-6710

IJIRSET

International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 11, Issue 3, March 2022

ISSN INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 7.569

9940 572 462

6381 907 438

ijirset@gmail.com

www.ijirset.com



Twitter Opinion Mining and Boosting Using Sentiment Analysis with SVR, RFA & RDT Algorithm

Pooja Dharasurkar¹, Gajanan Chakote²

P.G. Student, Department of Computer Science & Engineering, MSS's Engineering College, Jalna, Maharashtra, India¹

Assistant Professor, Department of Computer Science & Engineering, MSS's Engineering College, Jalna, Maharashtra, India²

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.

KEYWORDS: Support Vector Regression (SVR), Decision Trees (DTs), Twitter Sentiment Analysis.

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.

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.

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.

Twitter sentiment analysis allows you to keep track of what's being said about your product or service on social media, and can help you detect angry customers or negative mentions before they they escalate. 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. 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.

II. RELATED WORK

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.

Various studies that focus on analyzing social media data have been carried out, especially those related to specific events [1]. The intensive use of social media has attracted the attention of academic researchers, and many investigations have been conducted to obtain important information about the event. In recent years, substantial studies have been undertaken on sentiment analysis on Twitter. Jain and Dannavar [2] have investigated several steps for conducting sentiment analysis on Twitter data using machine learning algorithms. They also provide details of the proposed approach to sentiment analysis. This approach collects data by pre-processing data using the NLP technique. Then, feature extraction is carried out to get features that are relevant to the sentiment. Finally, a model is trained using classifiers, such as Naïve Bayes, Support Vector Machine (SVM), and Decision Tree. Another sentiment analysis was also carried out by Qianjun Shuai *et al.* [3] on the data of Chinese hotel reviews.

The research also using Doc2Vec is a feature for each document, then grouped the collected data into a balanced amount of positive and negative sentiment. After the Doc2Vec model is trained, the data is combined to being tested using classifiers such as SVM, Logistic Regression, and Naïve Bayes. From the results of this research, SVM gave the best results compared to other classifiers, with the F-Measure value of 81.16%

Sentiment analysis is the unification of various research fields such as NLP, data mining, and text mining with the aim of finding people's opinions expressed in written language [10]. There are several tasks in sentiment analysis such as sentiment extraction, sentiment classification, summarization, and many more [16]. This research will be a focus on sentiment classification. Sentiment can be classified as positive, negative, or neutral [16].

Sentiment analysis is not an easy thing to do. Several challenges in conducting sentiment analysis include synonym and polysemy, sarcasm, compound sentences, unstructured data, aspect, and many more [10]. For example, how the model can handle sentences with the same meaning but are described with different words (synonyms) or how the model can handle words with multiple meanings (polysemy). One way to handle this case is to use stemming. Stemming helps to process the word into its root word. this method cannot control the case if different word is used. These problems can be the cause of the resulting low performance.

Sentiment analysis is critical. Nowadays, people will very often post on social media such as Twitter. The data is very massive and unstructured. Therefore, it will be a challenge for humans to analyze the data and receive insight from the information effectively [3]. Sentiment analysis has come to solve this problem. Sentiment analysis can help organizations such as companies or even governments get some insight into people’s opinions about their decision.

III.METHODOLOGY

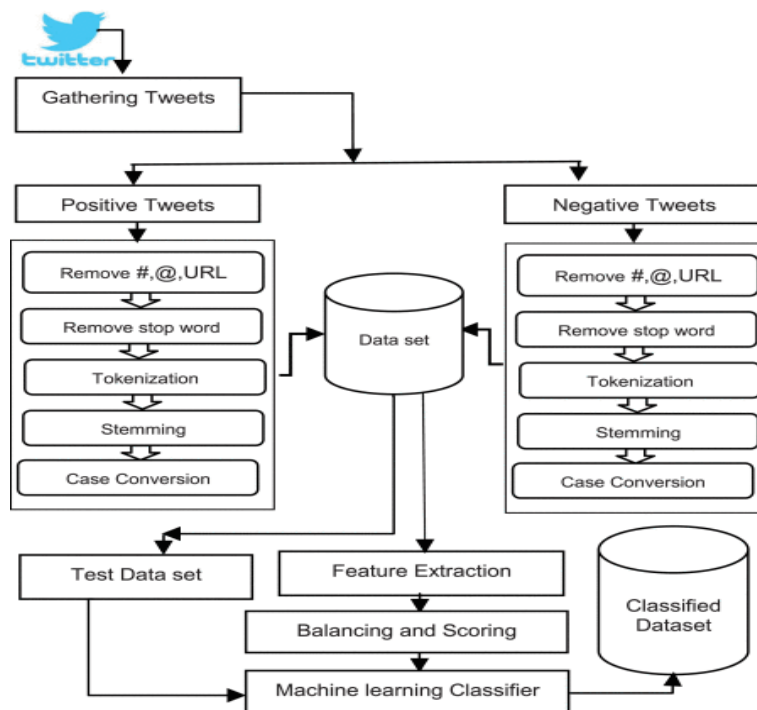


Fig.1.1 Twitter Opinion Mining and Boosting Using Sentiment Analysis

3.1 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.

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

3.2 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.

3.3. 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.

IV. EXPERIMENTAL RESULTS

4.1 Methods

4.1.1 Data collection

In this stage, tweet data is collected through crawling process using the Twitter API using the following words:

- Komodo
- Rinc
- Jurassic Park

Those words were chosen as a keyword because those words are very popular for this topic, like choosing komodo as a keyword because Rinca Island is the komodo habitat. Rinca is the name of the island where the development will be located. Jurassic Park because of the development in Pulau Rinca called Jurassic Park by Twitter users. The process produces 8000 tweet data that contain the keywords. The data collected through Twitter API is then removed from duplicate and unrelated data. After data cleansing, there are 1038 relevant tweet data.

4.1.2 Tagging

This process aims to train the model. Each tweet is tagged with a value of -1,0,1. Value 1 is a tweet that supports or is pro toward the development of Rinca Island. The value 0 is neutral with the development on Rinca Island, or the tweet is some news. The value -1 is a tweet that rejects or contradicts with development on Rinca Island.

The tagging process is done manually. Three people do a tag with the same tweets data. The final tagging value is the tagging value that is selected at most.

After Tagging all data, the distribution of each tagging can be seen in Table 1. The data in Table 1 is not balanced between labels. This can be seen from the number of labels worth 0 and 1, which has a composition of more than 75%.

4.1.3. Text preprocessing

At this stage, the tweet data that have been obtained are adjusted to suit the needs. Text preprocessing consists of 5 steps. The first step is cleansing. This process removes the data that only contain one or two words. After the data cleansing, the second step is to transform a word into its root word. The third steps are stop words and punctuation removal. The last process is tokenizing the sentence.

There are two alternatives to transforming a word into its root word, stemming and lemmatization. Lemmatization has better accuracy than stemming, as explained in the literature review, but stemming has better performance and not consuming as many resources as lemmatization.



In this research, stemming from the PySastrawi library was used. The stemming implemented Nazief Adriani Algorithm that claimed have 94.8% of accuracy for stemming in Bahasa [14]. The stemming process is not like normal stemming that only implements a stemming algorithm. After implementation of the stemming algorithm than the result is compared to their dictionary. This process is like the lemmatization process. Those processes are expected to get better accuracy in changing a word into its root word, than a normal stemming process. Stop word removal and punctuation removals such as punctuation or unrelated characters such as '@'.'#'. Last, tokenization aims to change a sentence or group of words into an array of words contained in the sentence. Those processes are also using PySastrawi Library. Pre-processing has a very important role in this research. The goal of all these processes is to get the data clean and ready to be processed. Clean data is expected to improve results in subsequent processes.

4.1.4. Data modelling

At this stage, the modeling results from the text that have been preprocessed were carried out using Doc2Vec. Preprocessed data will be separated into two groups. The first group is for training, and the second one is for testing.

The ratio between the two groups is 70:30. The vectors generated by Doc2Vec can be used to find similarities between sentences. To do data modeling will be using the Doc2Vec in Gensim Library. There are two types of models Doc2Vec used in this research: Distributed Memory Model of Paragraph Vectors (PV DM) and Distributed Bag of Words (PV- DBOW)

4.1.5. Sentiment classification

At this stage, sentiment classification is carried out on the data that has been modeled using SVM and Logistic Regression. The application of SVM and Logistic Regression is carried out by utilizing the Sklearn module and the Gensim library for the Doc2vec model for the distributed bag of word model and distributed memory for sentiment analysis processes the Python programming language.

4.1.6. Analysis and evaluation

This stage produces a sentiment classification of the tweet data and tests the result of the sentiment classification. Evaluation for this research is using several metrics such as accuracy and F-1 score.

4.2 Algorithms

4.2.1 Support Vector Regression

Support Vector Regression is a supervised learning algorithm that is used to predict discrete values. Support Vector Regression uses the same principle as the SVMs. The basic idea behind SVR is to find the best fit line. In SVR, the best fit line is the hyperplane that has the maximum number of points.

Step 1: Load the important libraries. ...

Step 2: Import dataset and extract the X variables and Y separately. ...

Step 3: Divide the dataset into train and test. ...

Step 4: Initializing the SVM classifier model. ...

Step 5: Fitting the SVM classifier model. ...

Step 6: Coming up with predictions.

4.2.2 Random Forest

Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase.

The Working process can be explained in the below steps and diagram:

Step-1: Select random K data points from the training set.

Step-2: Build the decision trees associated with the selected data points (Subsets).

Step-3: Choose the number N for decision trees that you want to build.

Step-4: Repeat Step 1 & 2.

4.2.3 Decision Tree Algorithm

1. Pick the best attribute/feature. The best attribute is one which best splits or separates the data.
2. Ask the relevant question.
3. Follow the answer path.
4. Go to step 1 until you arrive to the answer.

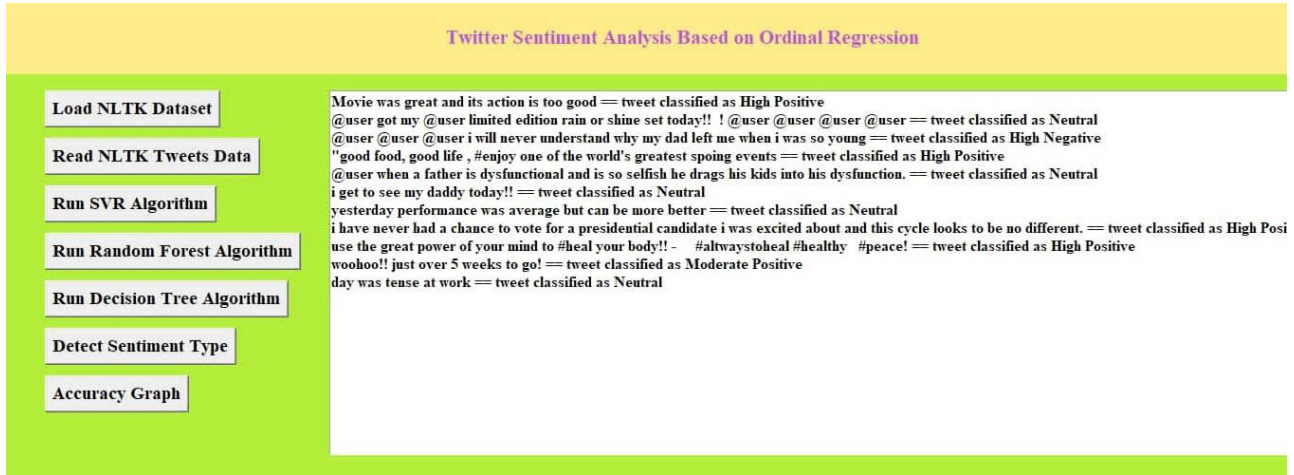


Fig 4.1 Twitter Sentiment Analysis Based on Ordinal Regression

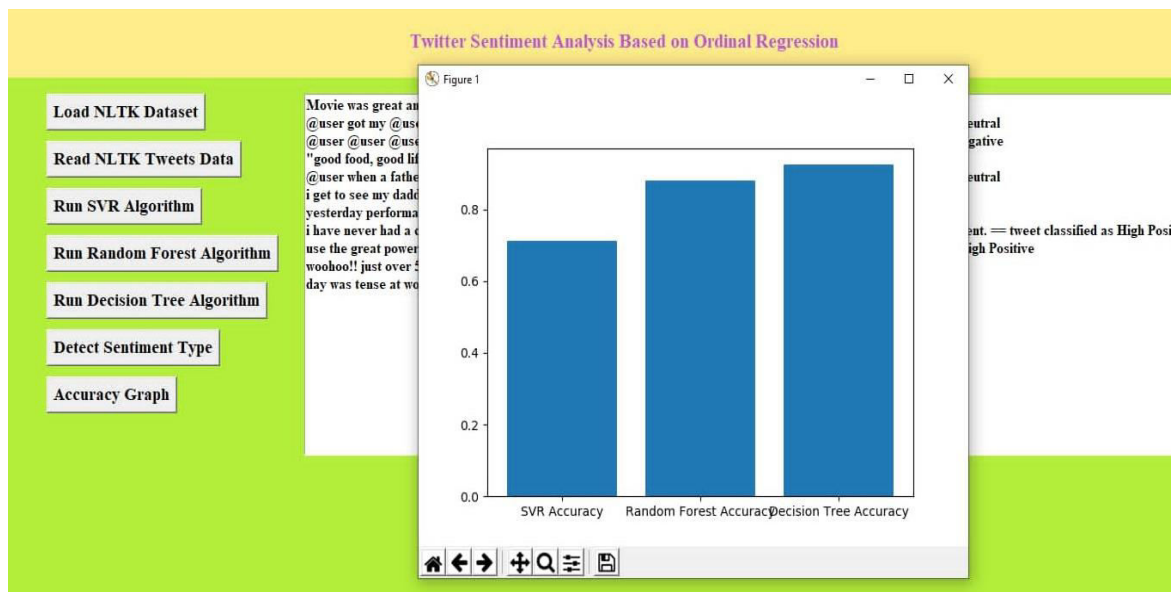


Fig 4.2 Comparison of Results with SVR, RFA & RDT Algorithm

V. 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 International 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.



**Impact Factor:
7.569**



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

 9940 572 462  6381 907 438  ijirset@gmail.com



www.ijirset.com

Scan to save the contact details