IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

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

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 10, Issue 9, September 2021

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.569

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 9, September 2021 ||

| DOI:10.15680/IJIRSET.2021.1009126 |

Twitter Opinion Mining and Boosting Using Sentiment Analysis

Pooja Dharasurkar [1], Gajanan Chakote [2]

P.G. Student, Department of Computer Science & Engineering, MSS's Engineering College, Jalna, Maharashtra, India[1]

Assistant Professor, Department of Computer Science & Engineering, MSS's Engineering College, Jalna, Maharashtra, India[2]

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 aAnalysis.

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.

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.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 9, September 2021 ||

| DOI:10.15680/IJIRSET.2021.1009126 |

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 specificevents [1]. The intensive use of social media has attracted the attention of academic researchers, and manyinvestigations have been conducted to obtain important information about the event. In recent years, substantial studieshave been undertaken on sentiment analysis on Twitter.Jain and Dannavar [2] have investigated several steps for conducting sentiment analysis on Twitter data usingmachine learning algorithms. They also provide details of the proposed approach to sentiment analysis. This approachcollects data by pre-processing data using the NLP technique. Then, feature extraction is carried out to get featuresthat 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 testedusing classifiers such as SVM, Logistic Regression, and Naïve Bayes. From the results of this research, SVM gavethe best results compared to other classifiers, with the F-Measure value of 81.16%

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 9, September 2021 ||

| DOI:10.15680/IJIRSET.2021.1009126 |

III. SYSTEM ARCHITECTURE

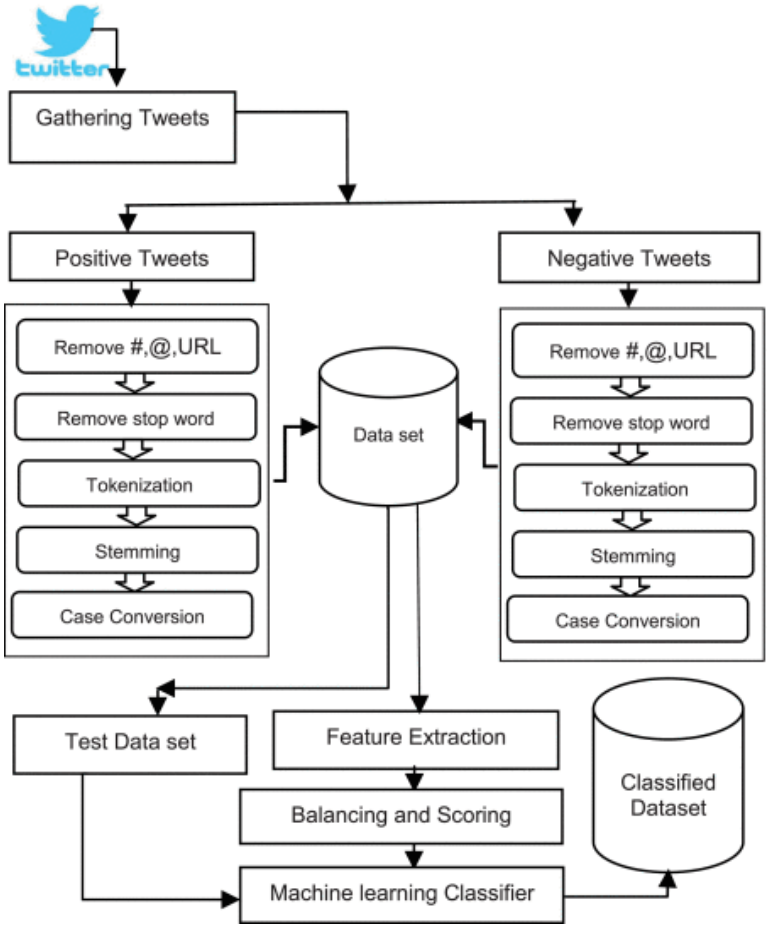

Fig.1 Twitter Opinion Mining and Boosting Using Sentiment Analysis

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");

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 9, September 2021 ||

| DOI:10.15680/IJIRSET.2021.1009126 |

- 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. 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.

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.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 9, September 2021 ||

| DOI:10.15680/IJIRSET.2021.1009126 |

[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 classi_cation 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: Areview," *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: Prac- tical 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 classi_cation," 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 speci_city 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.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.569

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

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