International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

REVIEW OF OPINION PREDIECTION BASED ON MACHINE LEARNING

BANDI FIDEL KASHTRO [1], Mrs.R. Madhuri Devi [2]

[1] PG Scholars, Department of CSE, ***PRIYADARSHINI INSTITUTE OF TECHNOLOGY AND MANAGEMENT***, Guntur Dist., Andhra Pradesh, India.

[2,] Associate Professor, Department of CSE, ***PRIYADARSHINI INSTITUTE OF TECHNOLOGY AND MANAGEMENT***, Guntur Dist., Andhra Pradesh, India.

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.

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

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

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.

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.LSTM has been widely used in NLP tasks. LSTM essentially has a memory like feature which allows it to better capture features, in the said task traditional methods are unyielding. Election prediction task intuitively requires a model to identify and weigh the features, same reasoning is used in natural language classification task where LSTMs are widely popular

LITERATURE SURVEY

Election prediction is a very popular problem amongst machine

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

learning community. In the past various approaches have been taken in order to solve the problem. Random forest and decision tree algorithms are among the popular ones. However, they have their own limitations and often fail to capture essential complex features involved in election prediction. A deep learning model would prove to be more successful in overcoming the limitations of traditional approaches. LSTM has been widely used in NLP tasks. LSTM essentially has a memory like feature which allows it to better capture features, in the said task traditional methods are unyielding. Election prediction task intuitively requires a model to identify and weigh the features, same reasoning is used in natural language classification task where LSTMs are widely popular.

EXISTING SYSTEM:

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. 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

Disadvantages

- Low performance
- Accuracy is less

PROPOSED SYSTEM

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

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

more effective as features.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.Li and Lin 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

Advantages

- High accurancy

METHODLOGY

3.1 METHODOLOGY

The method chosen for this research was a systematic literature review, which has proven to be a replicable and effective manner with which to identify, evaluate, interpret and compare studies that are relevant to a particular question or area [37]–[40]. The method used in this research followsthe guidelines defined by [40] and is fully described in Appendix I. This section presents the main points.

A.Research Questions

To define the research questions of this study, we returned to the main objectives: To provide a thorough review and investigation of the state of both the art and the practice of predicting election outcomes based on SM data and to identify key research challenges and opportunities in this field. Then, the following research questions were derived:

• RQ1: In which electoral contexts is the research being performed? This question aims at identifying the electoral contexts being studied, such as the year and country in which the election took place, and the type of election. This question is intended to ascertain whether the studies are best suited or giving attention to any particular electoral context.

• RQ2: What are the main approaches? The objective of this question is to identify the main approaches used, their main characteristics, how they

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

are modeled and applied to predict elections, and which are the metrics used to assess their performance.

• RQ3: What are the main characteristics of successful studies? The objective of this question is to identify the main characteristics of allegedly successful studies, in order to identify in which specific contexts, which approaches, and which factors yield effective results.

• RQ4. What are the main strengths and challenges of predicting elections with social media? After studying the context, approaches and characteristics of successful studies, the answer to this question aims to summarize the main perceived strengths, weaknesses, challenges, and opportunities in this new research area to guide future research.

B.Search Process

The rigor of the search process is one of the distinctive characteristics of systematic reviews [38]. To implement an unbiased and strict search, two approaches were combined: (i) automated search on indexing systems and (ii) snowballing search on the references of studies found on the automated search.

The automated search was performed in four indexing systems: ACM Digital Library, IEEEXplore Digital Library, ISI Web of Science, and Scopus. The search was performed on papers' metadata: title, abstract, and keywords and aimed to find studies focused on predicting elections based on SM data. Then, after some initial refinements, the following search string was used in the automatic search:

(model OR method OR approach OR framework) AND (predict*) AND (election*) AND ("social media" OR twitter OR facebook OR instagram).The snowballing search on the references was applied only at the end of study selection to perform this search only on already identified relevant studies.

C.Quality Assessment

v One initial difficulty regarding the quality assessment is thatthere is no established manner with which to define study "quality." In this study, we used the premise suggested by [37], in which quality relates to the extent to which the study minimizes bias and maximizes internal and external validity.Thus, we focused the quality assessment on the rigor of the study. Hence, we proposed the following quality assessment questions:

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

• QA1: Are the aim(s)/objective(s) clearly identified?

• QA2: Is the related work comprehensively reviewed?

• QA3: Are the findings/results clearly reported?

• QA4: Are bias and threats to validity clearly discussed?

• QA5: Did the study compare the proposed solution and results with other works?

Result:

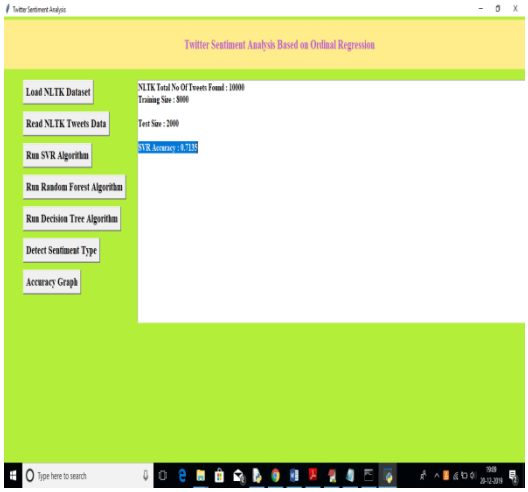
To run this project double click on 'run.bat' file to get below screen

In above screen click on 'Load NLTK Dataset' to load tweets dataset from NLTK library

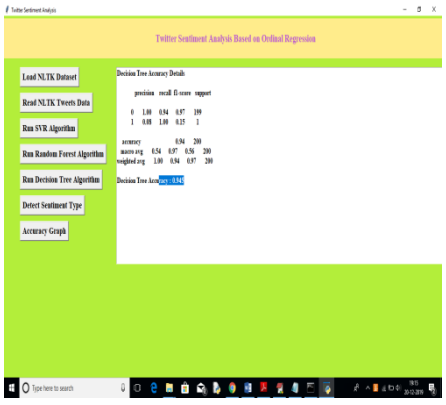
In above screen we can see total 10000 tweets are there in NLTK library, now click on 'Read NLTK Tweets Data' button to read all tweets and to build TFIDF vector. Upon each button click you need to wait for some seconds to get output. See below screen

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

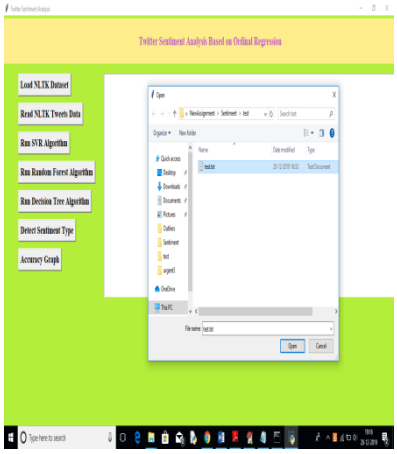
In above screen we can see total 8000 tweets vector used for training purpose and 2000 tweets used for testing purpose. Now click on 'Run SVR Algorithm' to build train model on that dataset and to calculate accuracy



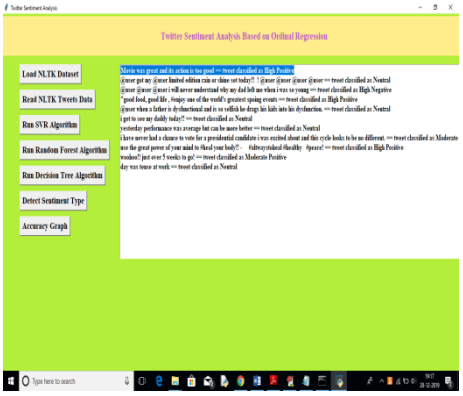
In above screen we can see SVR generate 0.71% prediction accuracy, now click on 'Run Random Forest Algorithm' button to calculate its accuracy



In above screen Decision Tree got 0.94% accuracy, Now click on 'Detect Sentiment Type' button and upload test tweets to predict it sentiment. In test folder inside test.txt you can see there is no sentiment label and application will detect I

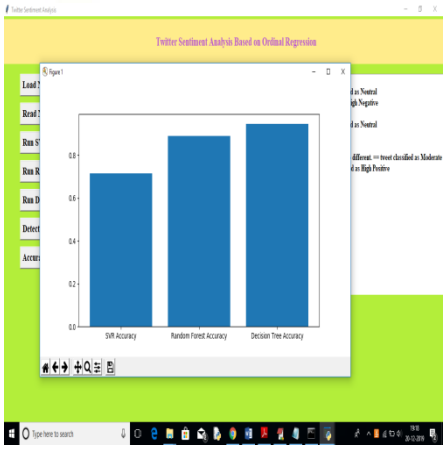


In above screen uploading test tweets file and below are the prediction results



In above screen for each tweet we can see the classified/predicted sentiments. Now click on 'Accuracy Button' to get below accuracy graph

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362



In above graph x-axis represents algorithm name and y-axis represents accuracy, from above graph we can see decision tree got better prediction compare to other algorithm.

In this paper author using another algorithm called SOFTMAX but its not a classifier algorithm, so I am not implementing it

CONCLUSION & FUTURE SCOPE:

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.

To allow a better evaluation of the studies' results, future works may focus on establishing a common framework of evaluation and common baselines. As was well discussed by, success must be measured statistically, not merely through description or mean average

error, and must be relative to clear benchmarks, which can be previous election results, existing polls, or default assumptions, such as incumbency success. Thus, the application of statistical tests, such as Wilcoxon signed-rank test, Wilcoxon–Mann–Whitney test, Welch's t-test, or paired t-test, just to cite a few, should be addressed. Finally, studies' reports should clearly discuss bias and threats to validity, together with the results

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 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: 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 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