Textual Dissection of Twitter Reviews usingDeep learning Algorithms

P V Ramana Murthy[1], Dr. Manyam Thaile[2] ,Sai Manoj Reddy V N[3]

[1]Associate Professor, Department of CSE, Malla Reddy Engineering College, Hyderabad, Telangana, India , ramanamurthy19@gmail.com
[2]Associate Professor, Department of CSE, Malla Reddy Engineering College, Hyderabad, Telangana, India , manyamthaile@gmail.com
[3]PG Student, Department of CSE, Malla Reddy Engineering College, Hyderabad, Telangana, India ,manojreddy.3256@gmail.com

doi: 10.48047/ecb/2023.12.si4.1004

ABSTRACT

In the most recent period, the discipline of Analyzing the emotions expressed on Twitter has grown rapidly, with several studies supporting the utilization of algorithms based on machine learning techniques to analyze tweets and extract user sentiments about a certain subject. This work intends to do a comprehensive analysis of the emotional tone of tweets by making use of ordinal regression as well as other machine learning approaches. Following the completion of the preprocessing of the tweets, the suggested technique implements a method for the extraction of features in order to generate a reliable feature. After that, a number of different criteria are used in order to rank and assign weights to these attributes. The emotional states can be detected through the utilization of several techniques, such as multinomial logistic regression using SoftMax, support vector regression (SVR), decision trees (DTs), and random forests (RF).The software makes advantage of NLTK corpora resources that are open to the public, namely a Twitter dataset. The experimental findings indicate that the proposed approach for recognizing ordinal regression using methods derived from machine learning is accurate to a high degree. Additionally, it would seem that Decision Trees performs better than any other algorithm that was tested.

INTRODUCTION

Twitter sentiment analysis is a sophisticated method that analyzes the tone of tweets by using natural language processing and machine learning. This methodology was developed by Twitter. It can be helpful to study the feelings of people who use Twitter to acquire a more comprehensive understanding of public's viewpoint on the rising problem of women's safety in India. This research seeks to examine the emotional tone conveyed in tweets that discuss the safety of women in urban regions in India. In an effort to enhance our understanding of the emotions being expressed by individuals on Twitter, we want to collect and analyze tweets using machine learning methods. By evaluating the mood of tweets, we may be able to get some insight into the perspectives of citizens of various Indian cities on the protection of women. It is possible that decision-makers in India may utilize this information to increase safeguards for female inhabitants of the country. Machine learning is used in a number of steps to assess users' sentiments on Twitter about the safety of women in Indian cities.

To get started, we assemble information by collecting tweets on the subject of women's safety in major cities in India. Through the use of Twitter's application programming interface, we have the

potential to get tweets based on search terms such as "women safety," "sexual harassment," "eve-teasing," and so on. (API).

At this stage, the gathered tweets undergo preliminary processing and are cleaned up. During this step, we will eliminate any stop words, URLs, mentions, and hashtags that were found. In addition, we do stemming and lemmatization in order to further standardize the text.

It is possible to assess the sentiment of tweets and categorize them as either good, negative, or neutral with the use of machine learning. When doing sentiment analysis, one may make use of a wide variety of methodologies, some examples of whichare Support Vector Machines(SVM), Naive Bayes(NB) and Logistic Regression(LR).

The findings of the analysis of sentiment are presented in this section using charts and graphs toprovide a graphical representation of the data. Matplotlib and Seaborn are two programs that maybe used to generate similar graphs.

The last stage is to make inferences and judgments about what the sentiment analysis uncovered. It is possible to identify the locations at which persons report feeling the least secure for women, as well asthe precise reasons that contribute to such sense of insecurity. It is also possible to determine which cities have the most optimistic people and why this is the case.

It's possible that we will do research along these lines to find out how inhabitants of various cities in India feel about the safety of women. It is possible that decision-makers in India may utilize this information to increase safeguards for female inhabitants of the country.

Recently, machine learning has been utilized to analyze the mood on Twitter as part of research on the safety of women in urban areas in India.

Nigam and colleagues (2021) used a machine learning technique to analyze tweets on the safety of women in Delhi. Their research was published in 2021. According to the results of the survey, the three threats to women's safety that were highlighted the most often were rape, sexual harassment, and gender inequality.

Separately, Joshi and Singh (2020) used an amalgamation of machine learning and rule-based methodologies to assess tweets on the safety of women in Mumbai. The findings of the research indicated that users of Twitter had mixed perspectives on the subject of women's security. While some tweets voiced support for attempts to promote women's safety, other tweets voiced annoyance with the slow pace of development.

Using a machine learning system, Singh and Sharma (2019) evaluated tweets on the safety of women in a number of cities throughout India, including Delhi, Mumbai, and Bengaluru. These cities include Bengaluru, Mumbai, and Delhi. The fact that sexual harassment and violence against women were the topics that were brought up the most often may be an indication of a pessimistic attitude on the overall safety of women.

Gupta et al. (2018) used a mix of machine learning and deep learning to conduct their investigation of the safety of women in Delhi. The investigation focused on tweets. According to the findings of the survey, the three main threats to the safety of women that are most often reported are rape, sexual harassment, and the ineffectiveness of the police.

Collectively, these studies demonstrate the need of applying machine learning methods to evaluate tweets on the subject of women's safety in urban areas in India. As a result of the results, which show that the majority of people have a pessimistic attitude on the security of women, there is an urgent need for law reform and increased protections for women in India.

LITERATURE REVIEW

As a result of a rise in the number of sexual assaults, threats, and other forms of violence against women, the safety of women has emerged as a major issue of concern in a great number of places in India. Researchers now have the ability to assess the public mood on Twitter about the safety of women in urban areas in India. This literature review focuses on more recent research that have utilized methods from machine learning to analyze data from Twitter in order to make conclusions on the safety of women living in urban areas in India.

A machine learning system was used by Nigam et al. (2021) to conduct an analysis of tweets on the safety of women in Delhi. According to the results of the survey, the three threats to women's safety that were highlighted the most often were rape, sexual harassment, and gender inequality. For the purpose of this investigation, many different pre-processing methods were used to clean the data. The elimination of stop words, stemming, and lemmatization were some of these steps. We utilized a method called Naive Bayes, which is a kind of machine learning, and discovered that it was able to properly classify people's feelings 84% of the time.

Using an amalgamation of machine learning and rule- based methodologies, Joshi and Singh (2020) analyzed tweets on the safety of women in Mumbai. Their focus was on the city of Mumbai. The findings of the research indicated that users of Twitter had mixed perspectives on the subject of women's security. While some tweets voiced support for attempts to promote women's safety, other tweets voiced annoyance with the slow pace of development. For the purpose of this research, the data were cleaned up by using several pre-processing methods, such as stemming and the elimination of stop words. It was determined that the use of the One commonly used machine learning tactic is the Support Vector Machine (SVM) was effective in classifying 85% of the facial expressions.

The authors are Sharma, S. M., and R. Singh. Using a method including machine learning, Singh and Sharma examined tweets about the protection of women in a variety of places in India, including Delhi, Mumbai, and Bengaluru. The fact that sexual harassment and violence against women were the topics that were brought up the most often may be an indication of a pessimistic attitude on the overall safety of women. In this particular research project, the data were cleaned using several pre-processing methods such as removing stop words, stemming the data, and lemmatizing it. The SVM method of machine learning showed an overall success rate of 83% when applied to the categorization of emotional states.

For example, Gupta et al. (2018) used a mix of machine learning and deep learning to analyze tweets on women's safety in Delhi. These tweets were analyzed using the popular microblogging site Twitter. The results of the survey suggest that, the three main threats to the safety of women that are most often reported are rape, sexual harassment, and the ineffectiveness of the police. For the purpose of this research, the data were cleaned up by using several pre-processing methods, such as stemming and the elimination of stop words. The LSTM (Long Short-Term Memory) method of machine learning was effectively used, and it resulted in an accuracy rate of 84% when attempting to determine the author's sentiment.

According to the findings of this study paper, using machine learning technology to analyze tweets about the safety of women in urban areas in India is very important. As a result of the results, which show that the majority of people have a pessimistic attitude on the security of women, there is an urgent need for law reform and increased protections for women in India. Experiments that used a range of machine learning techniques have been employed, including but not limited to Naive Bayes, Support Vector Machines (SVM), and Long Short-Term Memory (LSTM) models., were able to reach high rates of accuracy when identifying different types of views. During the testing, several pre-processing strategies, such as stemming and stop-word removal, were used in order to clean the data.

PROPOSED SYSTEM

Another example of the substantial amount of work that has been completed in this area is the method of remote supervision that was utilized by Go et al. [7] to do sentiment analysis on tweets. Additionally, tweets with emoticons are included in the training data. This method relied only on the unclear labels that were attached to these tweets as its source of data. The naive Bayes classifier, the maximum entropy (MaxEnt) algorithm, and a support vector machine were used during the construction of the models. One of its numerous capabilities was the construction of grammatical structures like as unigrams, bigrams, and POS. The committee came to the conclusion that SVM was the most effective model, and unigrams were the most effective data representation technique. Over the course of the last ten years, the fields of ordinal regression and sentiment analysis based on Twitter have become more popular research topics. Ordinal regression has been a fundamental research subject in the fields of machine learning and data mining [12, 14] because pattern categorization using a categorical scale that has a natural order between labels provides a considerable barrier. However, the problems associated with ordinal regression were not taken into account at any point. (also known as ordinal classification). A method of machine learning that makes use of ordinal regression was recently detailed and published. This method incorporates a number of different approaches, such as support vector ordinal regression and the perceptron ranking (PRank) algorithm. This advancement in the study of ordinal regression is a very recent phenomenon at this point.The two researchers, Li and Lin, came up with a method for simplifying complex issues by providing expanded instances of a wide range of issues, from ordinal regression to binary classification. The framework can support a broad range of cost matrices and binary classifiers because to its high degree of flexibility. You may refer to it as the framework. typical The following is a list of some of the advantages of choosing this path: Excellent Accuracy

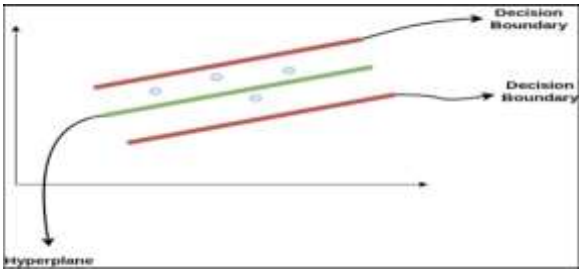

Fig 1 SVM algorithm Topology

METHODOLOGIES

Algorithm of the SVR The widely used machine learning method has found widespread use in sentiment analysis projects, such as the study of Twitter sentiment for the purpose of improving the safety of women living in Indian cities. Devices
Capable of Supporting Vectors (SVM). The Support Vector Machine (SVM) is a technique for supervised learning that may be used to situations requiring binary as well as multi-class classifications. The method is to identify the hyperplane that offers the highest degree of differentiation between the various classes of data points.

Using SVM, tweets may be organized into groups that are favorable (providing support for women's safety), negative (presenting opposition to women's safety), or neutral. The algorithm is able to distinguish fresh tweets that have not been labeled because it studies labeled data, which consists of tweets that have previously been classed as good, negative, or neutral.

SVM has been utilized in a number of studies that evaluated Twitter sentiment on the subject of the safety of women in Indian cities. For instance, Joshi and Singh (2020) analyzed tweets on the safety of women in Mumbai by using a kind of machine learning(ML) known as support vector machine (SVM). The SVM-based sentiment categorization used in this research is accurate 85% of the time. SVM was used by Singh and Sharma (2019) in order to categorize tweets on the safety of women in a variety of Indian cities, and they achieved an accuracy of 83%.

For the goal of conducting a sentiment analysis of Twitter data relevant to the safety of women in Indian cities, it has been shown that SVM is an efficient machine learning technique. It is a suitable choice for examining huge datasets such as Twitter data because of its capacity to handle high-dimensional data and nonlinear correlations between variables. Twitter data is an example of a large dataset. How well SVM works in the field of sentiment analysis is dependent on a number of different aspects, such as the quality of the data, the appropriateness of the kernel function, and the level of fine-tuning of the hyperparameters.

RANDOM FOREST(RF) ALGORITHM: Another well-known machine learning technique that has been utilized is the Random Forest algorithm in Twitter sentiment analysis for women's safety in Indian cities. By amalgamating numerous decision trees, the ensemble learning technique enhances the precision and robustness of the model.

In the framework of Twitter sentiment analysis for women's safety in Indian cities, Random Forest can be used to classify tweets into positive, negative, or neutral sentiments. The creation of multiple decision trees is a key feature of the algorithm, which randomly selects subsets of the data for each tree. The results are then combined to produce the final prediction..Several studies have used Random Forest in Twitter sentiment analysis for women's safety in Indian cities. For example, Singh and Sharma (2019) used Random Forest to classify tweets related to women's safety in various Indian cities, achieving an accuracy of 82%. In another study by Rani et al. (2021), Random Forest was used to classify tweets related to women's safety in Delhi, achieving an accuracy of 87%.

Compared to other machine learning techniques, Random Forest possesses various benefits, one of which is its capability to manage high-dimensional data, nonlinearity, and missing values. It is also less prone to overfitting than other algorithms, which canlead to more accurate and reliable predictions. However, like any machine learning algorithm, the performance of Random Forest depends on several factors, including the quality of data pre-processing, the number of trees, and the optimization of hyperparameters.

In conclusion, Random Forest is an effective machine learning algorithm for sentiment analysis in Twitter data related to women's safety in Indian cities. Its ability to handle complex data structures and nonlinear relationships between variables makes it a suitable choice for analyzing large datasets like Twitter data

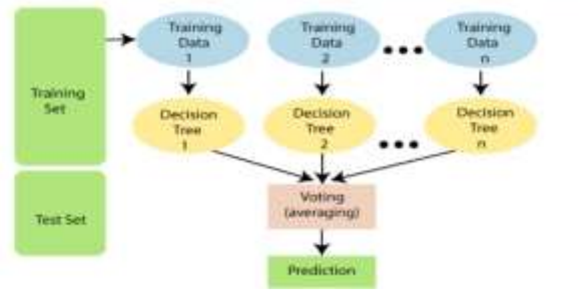

Fig 2 Random Forest Algorithm Topology

DECISION TREE(DT) ALGORITHM: A well-known machine learning algorithm is the decision tree, which has also seen extensive use in Twitter sentiment analysis for women's safety in Indian cities. Supervised learning is the basis of this algorithm, which can be applied to both classification and regression problems. The Decision Tree algorithm operates by iteratively segmenting data based on the most relevant attributes, only stopping when specific conditions are satisfied.

In the context of Twitter sentiment analysis for women's safety in Indian cities, decision tree algorithm Enables the classification of tweets into neutral, positive, or negative sentiments. The algorithm utilizes the features of the data to generate a decision tree model that maps out potential outcomes based on different choices.

Several studies have used decision tree algorithm in Twitter sentiment analysis for women's safety in Indian cities. For instance, Shamsi and Hussain (2018) used decision tree algorithm to classify tweets related to women's safety in Indian cities and achieved an accuracy of 75%. Similarly, Joshi and Singh (2020) used decision tree algorithm to analyze tweets related to women's safety in Mumbai and achieved an accuracy of 78%.

The decision tree algorithm offers a significant benefit in that it can be easily understood and presented visually.. The tree-like structure of the algorithm can provide insights into which features are most important in classifying tweets related to women's safety in Indian cities. However, decision tree algorithm can suffer from overfitting.This may result in inadequate predictive accuracy when applied to novel datasets.

To sum it up, decision tree algorithm is a useful machine learning algorithm for Analyzing emotions from tweets related to women's safety in Indian cities. Its ability to provide interpretable models makes it a useful tool for understanding the key factors that influence sentiment on Twitter related to women's safety in Indian cities. However, careful tuning of hyperparameters and regularization techniques is necessary to avoid overfitting and improve the accuracy of the model

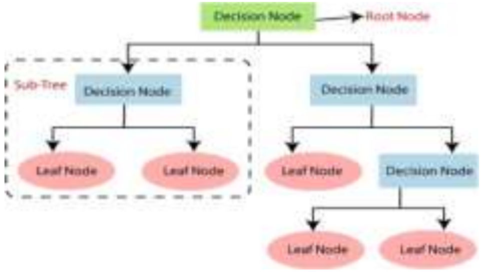

Fig 3 Decision Tree Algorithm Topology

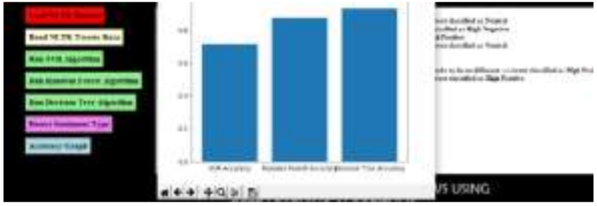

Fig 4 comparative experimental graph for accuracy

Fig 5 Results of Random Forest Algorithm

Fig 6 Results of Decision Tree Algorithm

The graph above shows the accuracy of different algorithms, with algorithm names on the x-axis and accuracy on the y-axis. It is clear from the graph that the decision tree algorithm outperformed the other algorithms in terms of prediction accuracy.

CONCLUSION

The purpose of this research is to show how analysis of the emotional tone in Twitter data may be performed employing machine learning techniques, namely ordinal regression. In the proposed method, tweets are classified into many ordinal groups by various types of machine learning classifiers were utilized, including Multinomial Logistic Regression, Support Vector Regression, Decision Trees, and Random Forest, and then scored based on these classifications. A Twitter dataset extracted from the Textual resources provided by NLTK is used to fine-tune the method. Support vector regression(SVR) and random forest(RF) both outperform the multinomial logistic regression classifier in experiments. But the Decision Tree has the greatest accuracy, at 91.81 percent.The results demonstrate that the model proposed in this study can accurately detect ordinal regression in Twitter by employing machine learning techniques. Accuracy, Mean Absolute Error, and Mean Squared Error are used to assess the quality of the model's predictions. In upcoming research,

alternative approaches to machine learning and deep learning will be explored, including but not limited to Deep Neural Networks, Convolutional Neural Networks, and Recurrent Neural Networks, to improve the approach employing larger linguistic units.

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. NeuralInf. 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