Journal of **N**onlinear **A**nalysis and **O**ptimization
Vol. 16, Issue. 1: 2025
ISSN : **1906-9685**

Twitter Sentiment Analysis: An NLP-Based Approach

#1 Dr. USHA SREE JAGARAGALLU, 2. VADLA VARSHINI, 3. THERAPOGU DHEERAJA RANI, 4. SAPURAM SHIRDHINI, 5. RATHLAVATH BRAMARI BAI
#1 Vice Principal in the Department of CSE, RAVINDRA COLLEGE OF ENGINEERING FOR WOMEN, KURNOOL.
#2#3#4 B. Tech in Computer Science and Engineering in RAVINDRA COLLEGE OF ENGINEERING FOR WOMEN, KURNOOL.

***Abstract*— Sentiment analysis on Twitter, which looks at tweets to find views, has grown rapidly in the last several years. Analysis using machine learning is common in the research community. Using machine learning and ordinal regression, this project will quantify the sentiment on Twitter. To improve efficiency, the approach collects properties from tweets and analyses them beforehand. Several courses offer score and balance. As far as sentiment analysis classification frameworks go, we suggest SoftMax, SVR, DTs, and RF. We built the approach using Twitter, a free NLTK corpus dataset. Experiments demonstrated that machine learning can accomplish ordinal regression detection. The results show that alternative solutions don't compare to Decision Trees.**

Keywords*— *Twitter, sentiment analysis, ordinal regression, ML

I. INTRODUCTION

The use of microblogging and social media has skyrocketed. One common way people express themselves online is through microblogging. Twitter generates a significant amount of data. Most of the current research has used social data to look at how people feel about certain products, issues, and occurrences. Often referred to as opinion mining, Natural language processing falls short when it comes to quantitative analysis. This method classifies the emotional tone of text as positive, negative, or neutral.Twitter sentiment analysis is a popular research tool. Big social data analysis is useful for collecting and categorising people's opinions. The unique characteristics of Twitter data make sentiment analysis more challenging compared to data from other sources. Acronyms and slang are acceptable in tweets due to the 140-character limit and the emphasis on informal discourse. To address these issues, researchers have investigated tweet sentiment analysis. The main approaches are lexicon-based sentiment analysis on Twitter and machine learning.The problem is ordinal categorisation, or regression. Ordinal regression is currently gaining popularity. A wide variety of disciplines investigate ordinal regression issues, which are essentially nominal problems with less-than-ideal solutions.Classification and regression issues are similar to ordinal regression problems, with a few key differences. This study analyses sentiment on Twitter by solving ordinal regression problems using machine learning. In order to classify tweets into categories, this study collects characteristics, creates a scoring and balancing system, and uses machine learning..

II. LITERATURE SURVEY

i) Statistical pattern recognition: a review

https://ieeexplore.ieee.org/document/824819

Classification, whether supervised or uncontrolled, is the primary emphasis of pattern recognition. There has been a lot of research and practical use of the statistical framework for pattern recognition. Methods based on statistical learning theory and neural networks have recently become popular. While planning a recognition system, it is important to think about things like pattern classes, sensing environments, pattern representation, feature extraction and selection, cluster analysis, classifier design and learning, training and test sample selection, and performance assessment. Recognising complicated patterns with varying orientation, location, and scale is a general difficulty that has not been solved after over 50 years of research and development. Data mining, web searching, multimedia data retrieval, face recognition, and cursive handwriting identification are just a few examples of the new applications that need rapid and robust pattern recognition methods. Known methods employed in pattern recognition system phases are summarised and compared in this review article, which also identifies state-of-the-art research topics and their potential applications.

ii) SemEval-2016 Task 4: Sentiment Analysis in Twitter

https://arxiv.org/abs/1912.01973

This report details the "Sentiment Analysis in Twitter Task" for its fourth year. Three of the five subtasks that

make up SemEval-2016 Task 4 are significantly distinct from one another. The first two objectives aim to predict the general and current tone of a tweet, much like in years past. Two variants of the core Twitter sentiment categorisation job are addressed by the three supplementary tasks. The initial version of the categorisation task is given an ordinal character by means of a five-point scale. Estimating the prevalence of each class of interest properly is what quantification in supervised learning literature is all about. Along with 43 other teams, this one is still quite popular.

iii) SemEval-2015 Task 10: Sentiment Analysis in Twitter

https://arxiv.org/abs/1912.02387

We cover the SemEval shared challenge on Twitter sentiment analysis that took place in 2015. Every year for the last three, more than forty teams took part in the most popular sentiment analysis shared assignment. This year's shared task competition included five subtasks related to sentiment prediction. (A) the sentiment of a phrase in a tweet and (B) the sentiment of a tweet were both carried over from previous years. We incorporated three new jobs to forecast (C) a tweet's emotion, (D) a sequence of tweets' sentiment as a whole, and (E) the phrase's previous polarity.

iv) Twitter Sentiment Classification using Distant Supervision

https://cs.stanford.edu/people/alecmgo/papers/TwitterDistantSupervision09.pdf

An unique approach is introduced for the purpose of automatically determining the sentiment on Twitter. Depending on the query word, messages are rated positively or negatively. People looking at product sentiment before purchase or companies keeping tabs on brand sentiment will find this tool useful. So yet, no studies have attempted to categorise the tone of tweets.

Here we offer the outcomes of machine learning systems that were supervised remotely and used to classify the sentiment of messages on Twitter. As a kind of noisy labelling, our training data consists of tweets that contain emoticons. It is possible to automatically get a large amount of training data from public sources. By using emoticon data, we show that SVM, Maximum Entropy, and Naive Bayes algorithms may reach 80% accuracy or higher.

The paper details the preprocessing steps that are essential for achieving high accuracy. An important takeaway from this research is the possibility of using emoticon-laden tweets for remote supervised learning.

v) Application of machine learning techniques to sentiment analysis

https://ieeexplore.ieee.org/document/7912076

The 'data age' has arrived. Twitter and other user-generated content platforms have opened up new opportunities for businesses that monitor product reviews and comments. Politics, goods, and sports may all be found on the rapidly expanding microblogging site known as Twitter. Everyone from businesses to governments to people stands to gain from these views. One way to gauge public sentiment is through the usage of tweets. Whether it's about a product, person, issue, event, etc., sentiment analysis can automatically tell if user-generated content is positive, negative, or neutral. Twitter sentiment analysis using machine learning is described in this article. This paper also describes the sentiment analysis approach. This article presents a flexible, fast, and scalable infrastructure for Twitter text analysis that is built on Apache Spark. The machine learning techniques of Naïve Bayes and Decision Trees are utilised in the proposed framework for sentiment analysis.

III. METHODOLOGY

A. *Proposed Work:*

Four essential components constitute the proposed system. To facilitate sentiment analysis, the first module gathers tagged tweets; the second module preprocesses them. To build a classification model, Module 3 extracts important characteristics. Next, we display the score and balance of tweets. In the last module, tweets are categorised as either very positive, moderately positive, neutral, moderately negative, or highly negative. This makes use of machine learning classifiers. A representation of the sentiment analysis process by a machine learning algorithm is shown in Figure 1. The proposed model can detect ordinal regression in Twitter quite well. Accuracy, Mean Absolute Error, and Mean Squared Error are metrics used to measure the performance of a model..

B. *System Architecture:*

People can easily ascertain the quality of any business or product through topical opinion mining on the internet. Users will provide positive evaluations and thoughts when they have a positive encounter with a product or company. This feedback helps other users to judge the product's quality. Everybody says what they think on modern social media platforms like Twitter, and these platforms constantly innovate new ways to tease out underlying feelings in user-generated content. As a means of attitude detection, the author employs Random Forest, Support Vector Regression, Decision Tree, and Softmax..

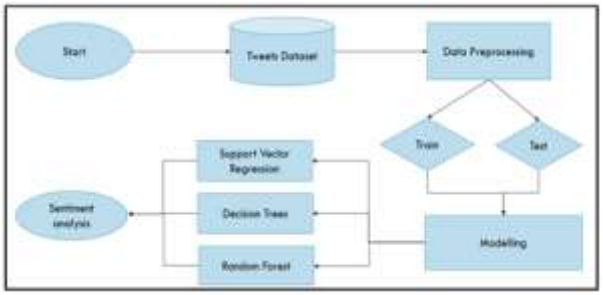

Fig.1. Proposed architecture

C. *Dataset Collection:*

We use the publicly available NLTK Corpora Resources dataset from Twitter. Experimental findings show that the proposed method can correctly detect ordinal regression using machine learning techniques. When compared to other algorithms, decision trees consistently deliver the best results. In Figure 1, we see the sentiment analysis method in action. After cleaning and splitting the dataset into training and testing sets, we will preprocess it using the following methods. We will extract certain traits from the dataset to narrow it down. Please create a decision tree model to

determine if a tweet is excellent or terrible next. The decision tree will be retested using real-time tweets..

	Apple	day	keep	doctor	away	good health	(unique words from 2 tweets)
T1_1	1	1	1	1	0	0	(count of all words from tweet1)
T2_2	0	0	0	0	1	1	

Fig.2. Dataset

D. Data Processing:

We gathered the initial data. The best approach to guarantee correctness, predictability, and usefulness is to cleanse the data. It is common for datasets to need cleaning because of outliers, which are pieces of data that are either undesired or very noisy. Unexpected outcomes might result from outliers. A more trustworthy and consistent dataset is the end result of data cleaning, which involves removing and editing unnecessary data. Here are several approaches for cleaning data. By using this module, NLTK tweets may be parsed, with special characters removed, words halted, and stemmed (for instance, ORGANISATION becomes ORGANISE). Afterwards, the TFIDF vector is computed..

E. Feature Selection:

To train a classification model, feature extraction is used to get relevant attributes. You may utilise our recovered features with raw data in various formats, such as text, symbols, and images, and they are well-suited for machine learning techniques. Count vectorisation and TF-IDF are among the techniques used to extract features from tweets. In order to find the feature matrix that indicates the importance of phrases in a text corpus, this study used TF-IDF.

V. IMPLEMENTATION

1. MODULES:

i) Load NLTK Tweets:

The NLTK Twitter sentiment corpora are loaded by this module.

ii) Read NLTK Tweets:

By using this module, NLTK tweets may be parsed, with special characters removed, words halted, and stemmed. Afterwards, the TFIDF vector is computed.

iii) Run SVR Algorithm:

In this module we will give TFIDF vector as input to train SVR algorithm. This algorithm will take 80% vector for train and 20% vector as test. Then algorithm applied 80% trained model on 20% test data to calculate prediction accuracy.

iv) Similarly we will build model for Random Forest and Decision tree to calculate their accuracy.

v) Detect Sentiment Type:

Here we may use a train model to forecast sentiment based on test tweets.

vi) *Accuracy Graph:* With this module, you may see graphs showing the correctness of all your algorithms.

2. ALGORITHMS:

a) Ordinal Regression:

The ordinal regression method predicts the data class using a number of independent variables. We use tweets as input and train a classifier to predict sentiment based on all the independent keywords in this study. By analysing independent variables, ordinal regression may forecast ordinal dependent variables. Both continuous and categorical variables can be used in this study; the dependent variable is the order of the answers.

b) Softmax:

Although cumulative logit or ordered logistic regression can modify softmax for ordinal regression, its primary use is in multiclass classification. Instead of treating the categories as unordered, this change takes their underlying order into consideration. Predictions made with this approach maintain class ordinal relationships.

c) Decision Tree:

Altering the splitting criterion for the order of the target variables allows for the modification of decision trees for ordinal regression. The tree may be trained to reduce target-ordered errors, such as mean squared error, rather than impurity-related errors, such as the Gini index or entropy. This aids the model in understanding sequences of ordinal variables.

d) Random Forest:

Ordinal regression may make use of random forests, which are collections of decision trees. The last forecast is often the most prevalent class, and every tree in the forest predicts an ordinal value. In order to train the model to perform better on ordinal tasks, we may optimise loss functions that take the intended class ordering into consideration.

e) Support Vector Regression (SVR):

Support By treating ordinal labels as continuous values and using loss functions that punish predictions based on their distance from the actual ordinal value, vector regression may be used to do ordinal regression. Ordinal data with sequential but non-linear correlations can be fit using the model, which preserves class order while fitting a regression line.

IV. EXPERIMENTAL RESULTS

Our training of all algorithms using the publicly available Twitter dataset from the NLTK toolkit yielded better predictions than Decision Tree. We sanitise Twitter material using NLTK (Natural Language Processing Tool Kit) library capabilities, such as deleting word stems (ing, tion, etc.), special symbols, and stop words (the, then, there). Once the tweets have been cleaned, we will transform them into BOG (Bag of Words Dictionary). Then, we will use TF/IDF to transform BOG into a vector.

TF/IDF = number of times word occur in tweet / total number of times word occur in all tweets.

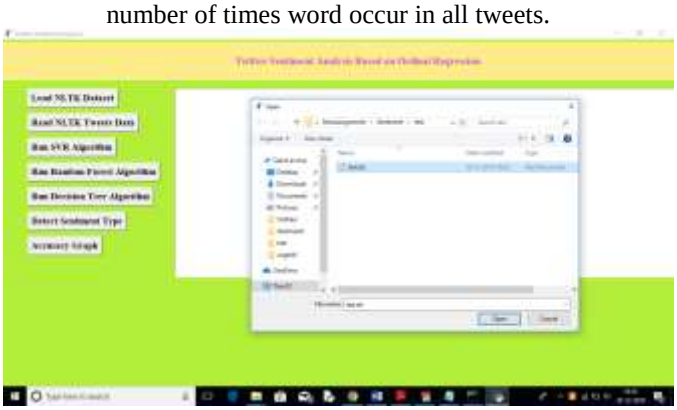

Fig.3. Dataset load

Fig.4. classified/predicted sentiments

Fig.5. Graph

VI. CONCLUSION

For ordinal regression, machine learning is utilised to evaluate sentiment on Twitter. We begin by developing a model for Twitter sentiment analysis that includes scoring and balancing. Afterwards, we use machine learning classifiers to sort tweets into ordinal groups. A few examples of research classifiers include Support Vector Regression, MultinomialLR, DT, and RF. Using NLTK corpora, this method optimises on publicly available Twitter data.

According to the results, Random Forest and Support Vector Regression are more accurate than Multinomial Logistic Regression. The Decision Tree does a good job with 91.81 percent accuracy. The proposed machine learning technique detects ordinal regression in Twitter, as shown in trials. The precision, MSE, and MAE are the metrics by which the model's efficacy is evaluated.

VII. FUTURE SCOPE

Enhance our approach by making use of bigrams and trigrams. We want to investigate various machine learning and deep learning techniques, including convolutional, recurrent, and deep neural networks.

REFERENCES

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

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

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

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

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

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

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

[8] M. Bouazizi and T. Ohtsuki, ''A pattern-based approach for multi-class sentiment analysis in Twitter,'' IEEE Access, vol. 5, pp. 20617–20639, 2017.

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

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

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

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

Author' s Profile

➢ Dr. **Usha Sree Jagaragallu** is a **Professor** and the **Vice Principal** in the Department of **Computer Science & Engineering** at **Ravindra College of Engineering for Women**, **Kurnool**.

Email: viceprincipal@recw.ac.in

➢ I am **Vadla Varshini**, currently pursuing a B.Tech in Computer Science and Engineering at **Ravindra College of Engineering for Women**, Kurnool, Andhra Pradesh. I have a strong passion for technology and programming, with key areas of interest including **Java, Python, HTML, CSS, and JavaScript**.

As part of my learning journey, I have successfully completed an **internship at Coincent.ai**, where I worked on **Artificial Intelligence projects using Python**. During this internship, I gained valuable hands-on experience in building intelligent systems, understanding machine learning algorithms, and developing AI-driven solutions. This experience has significantly enhanced my problem-solving skills and deepened

my understanding of how AI can be applied in real-world scenarios.

➢ I am **Therapogu Dheeraja Rani**, currently pursuing a **B.Tech in Computer Science and Engineering** at **Ravindra College of Engineering for Women**, **Kurnool, Andhra Pradesh, India**. I have a keen interest in the field of technology, particularly in **Java, Python, HTML, and CSS**. I am passionate about learning and continuously enhancing my programming skills to build **innovative and efficient solutions**. To strengthen my foundation in **Data Structures and Algorithms**, I have been certified as a **Smart Coder** from **Smart Interviews**. Additionally, I earned a certification in **Web Development** from **Coincent**, which enabled me to explore and implement real-world web technologies.

One of my recent projects, **Website Creation for a Car Dealership**, involved designing and developing a **dynamic and visually appealing website** using **HTML and CSS**. The goal of the project was to create a **digital showroom** that allows users to explore various car models, view **detailed specifications**, and access essential **dealership information**. I structured the content across multiple pages, including the **Homepage**, **Car Listings**, **Car Details**, and a **Contact Page**, using **HTML**. **CSS** was used extensively to **style the website**, **incorporate images**, and ensure a **user-friendly and attractive interface**. This hands-on experience enhanced my **front-end development skills** and deepened my understanding of building **responsive web interfaces**.

➢ I am **Sapuram Shirdhini**, currently pursuing a Bachelor of Technology (B.Tech) in Computer Science and Engineering at **Ravindra College of Engineering for Women**, located in Kurnool, Andhra Pradesh, India. My technical interests lie in **Java, HTML, CSS, and SQL**, and I am passionate about exploring the field of web development and design.

As part of my learning journey, I worked as a **Web Development and Designing Intern at Oasis Infobyte**, where I had the opportunity to enhance my front-end development skills. During the internship, I developed a **Temperature Converter** application using **HTML, CSS, and JavaScript**. The project focused on creating a user-friendly interface that allows users to easily convert temperatures between **Celsius, Fahrenheit, and Kelvin**. This experience not only strengthened my technical skills but also gave me valuable insights into building responsive and interactive web applications.

➢ I am **Rathlavath Bramari Bai**, currently pursuing a **Bachelor of Technology in Computer Science and Engineering** at **Ravindra College of Engineering for Women**, located in **Kurnool, Andhra Pradesh, India**. I have a keen interest in technologies such as **Python, Java, HTML, CSS, SQL**, and **Cloud Computing**, and I'm passionate about building a strong foundation in both programming and emerging tech domains.

To enhance my skills and gain industry-relevant knowledge, I have completed **"Data Science for Beginners"** offered by **Board Infinity** and a course on **Artificial Intelligence** by **Coincent**. These learning experiences have helped me develop a deeper understanding of data-driven technologies and AI concepts, fueling my motivation to explore innovative solutions in the tech world.