IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

TWITTER SENTIMENT ANALYSIS BASED ON ORDINAL REGRESSION

[1]Dr.A Swathi, [2] J Sri Sai Priyatha, [3]K Akshitha Ravinder, [4]G Adithya Goud, [5]K Prashanth.
[1]Associate Professor, [2,3,4,5] Student,
[1]Sreyas Institute of Engineering and Technology, Hyderabad, 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.

Index Terms - Support Vector Regression (SVR), Decision Trees (DTs), Random Forest (RF), Twitter Dataset, NLTK Corpora Resources.

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. 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. Twitter sentiment analysis approaches can be generally categorized into two main approaches, the machine learning approach, and a lexicon-based approach.

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.

Ordinal regression problems with some similarities and differences can be said to be between classification and regression. 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. LITERATURE SURVEY:

[1] Palomino et al.in "Evaluating the Effectiveness of Text Pre-Processing in Sentiment Analysis" in 2022 introduced in their research about the critical evaluation of text pre-processing techniques in the domain of sentiment analysis, which has garnered immense attention due to both practical demands and academic interests. With social media serving as a primary source for sentiment analysis and the text within these platforms often deviating from standard grammatical rules, the need for pre-processing techniques to cleanse and normalize data becomes imperative. While pre-processing methods have been extensively discussed in the literature, a definitive consensus on best practices remains elusive. The study systematically reviews existing research and quantitatively evaluates various combinations of pre-processing components. Focusing on Twitter sentiment analysis, acknowledged as a significant public data source, the research assesses the efficacy of different pre-processing combinations on the overall accuracy of off-the-shelf tools and an algorithm developed by the authors. Their findings highlight the significance of the order of pre-processing components, showcasing a substantial enhancement in the performance of naïve Bayes classifiers. Additionally, the study reveals that while lemmatization contributes to improving an index's performance, its impact on the quality of sentiment analysis remains less significant. This comprehensive study significantly contributes to the field by shedding light on the nuanced impacts of different pre-processing techniques, aiding in the refinement of sentiment analysis methodologies, especially within the realm of Twitter data analysis.
[2] Hassan et al. (2021) explored an innovative approach in their study "Altmetrics: A machine learning approach" aimed at annotating an Altmetrics dataset from diverse disciplines and conducting sentiment analysis using machine learning and natural language processing algorithms. The research team initially provided guidelines to human annotators, resulting in an inter-annotator agreement (IAA) of 0.80 (Cohen's Kappa) after appropriately labeling the sentiments in tweets related to scientific literature. Evaluations were conducted on two versions of the dataset: one comprising English tweets and the other including 23 languages, among which English was prominent. Leveraging 6388 tweets about 300 papers indexed in Web of Science, the study compared the efficiency of various machine learning models with established sentiment analysis benchmarks like SentiStrength and Sentiment140. Their findings showcased that the Support Vector Machine utilizing uni-gram surpassed all classifiers and baselines, achieving an accuracy exceeding 85%. Additionally, Logistic Regression demonstrated an 83% accuracy, while Naïve Bayes achieved an 80% accuracy. The precision, recall, and F1 scores for Support Vector Machine, Logistic Regression, and Naïve Bayes were notably high, providing (0.89, 0.86, 0.86), (0.86, 0.83, 0.80), and (0.85, 0.81, 0.76) respectively, underlining the efficacy of their proposed models in sentiment analysis of Altmetrics datasets. This study not only identified the most effective model but also offered a Python library for sentiment analysis, contributing significantly to the field of sentiment analysis for scholarly use cases.
[3] Yadav et al. explored an innovative approach in their study "Twitter Sentiment Analysis Using Machine Learning For Product Evaluation" in 2020 introduced in their research about utilizing Twitter as a rich source of public opinions, providing a massive repository for sentiments expressed towards a wide array of subjects, including people, services, companies, and products. This paper centers on sentiment analysis, a process crucial for analyzing public viewpoints, particularly when merged with Twitter data, which offers valuable insights into the sentiments expressed on the platform. Recognizing the significance of social media as an invaluable resource for understanding user preferences and feedback, the study underscores the importance of sentiment analysis in assessing public opinions towards specific products or services. The research highlights various techniques employed for classifying product reviews, often presented in the form of tweets, to discern whether the sentiments expressed are positive, negative, or neutral. Emphasizing the significance of this analysis in evaluating the product market, the study utilizes online product reviews gathered from Twitter to determine the most effective classifier for sentiment analysis. This work significantly contributes to understanding the methodologies and their implications for product evaluation in the context of sentiment analysis using Twitter data, providing insights into the process of assessing public opinions and their relevance for market evaluation.
[4] Kariya et al. explored an innovative approach in their study "Twitter Sentiment Analysis" in their study introduced about the realm of Twitter sentiment analysis in the context of the burgeoning social media landscape. The paper focuses on the extraction of sentiments from tweets, presenting an approach to classify tweets into positive, negative, or neutral categories. Recognizing the potential utility of this approach for organizations mentioned or tagged in tweets, the study underscores the unstructured nature of tweets,

emphasizing the initial requirement to convert them into a structured format. The research details a pre-processing phase to resolve the unstructured tweet format and outlines the utilization of Twitter API libraries for tweet access. A crucial aspect of their work involves training datasets using algorithms to enable the categorization of tweets, thereby facilitating the extraction of relevant sentiments from the analyzed content. This study significantly contributes to the field of sentiment analysis, offering insights and methodologies for structuring and analyzing unstructured tweets, ultimately aiding organizations in gleaning valuable sentiments expressed on Twitter.

[5] Mandloi et al. explored an innovative approach in their study "Twitter Sentiments Analysis Using Machine Learning Methods" in 2020 researched on the critical task of sentiment analysis, which involves discerning whether text conveys positive, negative, or neutral sentiments, also known as material polarity or opinion mining. Their work acknowledges the significant growth and evolution of social media platforms, particularly emphasizing Twitter, where users share succinct messages of 280 characters. The limited length of tweets simplifies sentiment analysis. Twitter, being a platform with a colossal user base posting approximately 550 million tweets daily, offers a vast and diverse dataset reflecting various age groups and gender representations, making it an excellent source for societal sentiment analysis. The paper aims to compare multiple machine learning methods such as Naïve Bayes Classification, Support Vector Machine Classification, and Maximum Entropy Classification for sentiment analysis. Their investigation focuses on understanding how these classification algorithms perform sentiment analysis and evaluates their accuracy and precision in such tasks. This research contributes significantly to understanding the efficacy of different machine learning methods in analyzing sentiments expressed on Twitter, providing insights into their comparative performance and potential for societal sentiment analysis.

[6] Elbagir et al. explored an innovative approach in their study "Twitter Sentiment Analysis Using Natural Language Toolkit and VADER Sentiment" in 2019 conducted a significant study focusing on the sentiment analysis of Twitter data, which has emerged as a prominent area of research. Their investigation aimed to leverage the Valence Aware Dictionary for sentiment Reasoner, (VADER) for the classification of sentiments expressed within Twitter data. Notably, the research departed from previous studies primarily oriented towards binary classification and instead introduced a multi-classification system for analysing tweets. Specifically, they applied VADER to classify tweets associated with the 2016 US election. Their findings revealed promising results, demonstrating strong accuracy in detecting ternary and multiple sentiment classes, indicating the effectiveness of VADER in handling more nuanced sentiment analysis within the context of Twitter data. This study contributes significantly by expanding the scope of sentiment analysis beyond binary classifications, offering valuable insights into the complex nature of sentiment expressions in social media discourse.

[7] Wagh et al. in their study "Survey on Sentiment Analysis using Twitter Dataset" in 2018 published a survey paper that delves into the realm of sentiment analysis using Twitter datasets, acknowledging the vast number of individuals expressing their thoughts through tweets on social networking sites like Twitter. Recognizing the brevity and simplicity inherent in tweets as a basic means of expression, the study focuses on sentiment analysis as a component of text data mining and natural language processing (NLP). It highlights the diverse approaches and techniques employed in sentiment analysis for Twitter data. The paper categorizes sentiment analysis types and details various methods used for sentiment extraction from tweets. Engaging in a comparative study, the research explores different techniques and approaches in sentiment analysis specifically concerning Twitter data. This survey paper significantly contributes to the comprehensive understanding of sentiment analysis methodologies used in the context of Twitter, providing insights into various approaches and techniques, thereby serving as a valuable resource for researchers and practitioners in the field of sentiment analysis using social media data.

[8] Hasan et al. explored an innovative approach in their study "Machine Learning-Based Sentiment Analysis for Twitter Accounts (2018)" in their study introduced the burgeoning field of opinion mining and sentiment analysis, focusing on the evaluation of opinions and text present across various social media platforms using machine learning methodologies. Their research, accepted in February 2018, addresses the swift growth in sentiment analysis techniques, particularly in the context of political opinions expressed during elections. Acknowledging the need for more advanced methodologies, the paper introduces a hybrid approach involving a sentiment analyzer employing machine learning techniques. A significant contribution of their work lies in providing a comparative analysis of sentiment analysis techniques, specifically in the realm of political views. They apply supervised machine learning algorithms such as Naïve Bayes and support vector machines (SVM) to analyze and categorize political sentiments. This research offers insights into the efficiency of different sentiment analysis methods, particularly within the context of political discourse on social media, thereby contributing to the advancement of sentiment analysis in the domain of public opinion during elections.

[9] Bhumika Gupta et al. introduced “Twitter Sentiment Analysis using machine learning algorithms in Python”. Their research focuses on the analysis of sentiments expressed on Twitter, aiming to extract and understand the emotions and opinions conveyed by users. Twitter, as a platform, presents a challenge due to the concise and informal nature of tweets, often laden with slang, abbreviations, and varied expressions. The team's literature survey reviews previous works in sentiment analysis on Twitter, highlighting the methodologies, models, and approaches adopted by various researchers. The objective is to present a comprehensive overview of the field, culminating in a generalized Python-based approach. This research is significant given the consistent growth and evolving challenges posed by the unique format of Twitter data, providing valuable insights into sentiment analysis techniques for social media content.
[10] Krouska et al. in their work “The effect of preprocessing techniques on Twitter Sentiment Analysis” in 2016 introduced in their research about the essential role of data preprocessing in Twitter sentiment analysis, acknowledging Twitter as a valuable platform for expressing diverse thoughts and opinions. The paper highlights the significance of properly identified reviews as a fundamental input for various systems, including e-learning and decision support systems. Emphasizing the critical nature of data preprocessing in sentiment analysis, the study explores how selecting appropriate preprocessing methods can significantly enhance the accurate classification of instances. The research details the essential procedures for preprocessing reviews to identify sentiment and perform analysis, specifically in determining whether the sentiment expressed is positive or negative. The paper provides an extensive comparison of sentiment polarity classification methods tailored for Twitter text and deeply investigates the role of text preprocessing in sentiment analysis. Through a series of tests involving various combinations of methods and experiments conducted on manually annotated Twitter datasets, the research highlights the impact of feature selection and representation on positively influencing classification performance. This study serves as a valuable resource, shedding light on the crucial aspects of data preprocessing in Twitter sentiment analysis, offering insights and methodologies that can significantly enhance sentiment classification accuracy. [11-17] implemented various deep learning methods to predict age of a person using different activation functions and optimizers to predict the accuracy.

Table 1

Sl No:	**Author**	**Year of Publication**	**Proposed System/Algorithm**	**Results/Comment**
1	Marco A. Palomino	2022	Naïve Bayes Classifier algorithm	86%
2	*Hassan*	2021	Support Vector Machine (SVM), Logistic Regression, Naïve Bayes	90%
3	Akrivi Krouska	2020	Algorithms like Naïve Bayes, SVM, KNN, and C4.5 were used to analyze sentiments from Twitter data (datasets on Obama-McCain Debate, Health Care Reform, and Stanford Twitter Sentiment Gold Standard), achieving varying accuracy depending on different preprocessing methods like unigrams, bigrams, and 1-3 grams.	Unigram: NB: 88.08% (Top 70% attribute selection), 88.25% (InfoGain attribute selection) Bigram: NB: 89.02% (InfoGain attribute selection), 87.52% (Top 70% attribute selection) 1-3 grams: NB: 91.59% (InfoGain attribute selection), 91.94% (Top 70% attribute selection)

4	Nikhil Yadav	2020	XGBoost	XGBoost- 70-80%
5	Chirag Kariya	2020	KNN	90.912%
6	Lokesh Mandloi	2020	Naïve Bayes Classifier	86%
7	Shihab Elbagir	2019	NLTK and the VADER analyzer were applied to conduct a sentiment analysis	56%
8	Rasika Wagh	2018	NaviveBayes , SVM	Naïve Bayes-85.5% SVM-90%
9	Ali Hasan	2018	TextBlob, SentiWordNet, W-WSD	TextBlob – 76% SentiWordNet – 54% W-WSD- 79%
10	Bhumika et al.[]	2017	Maximum Entropy, Ensemble classifier	MaximumEntropy- 90.0% Ensemble classifier- 90.0%

III. METHODOLOGY:

The proposed system involves the following steps: data collecting, cleaning, preprocessing, and applying the model to obtain the desired output.

3.1 Data Collection

The 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. Figure 1 depicts the overall process of sentiment analysis. The dataset will be preprocessed using the following methods once cleaned and partitioned (separated) into training and testing datasets. To make the dataset smaller, features will be extracted. The next step is to develop a model that the decision tree will use to categorise the tweets into positive and negative ones. Once more, the decision tree will get real-time tweets to test the real-time data.

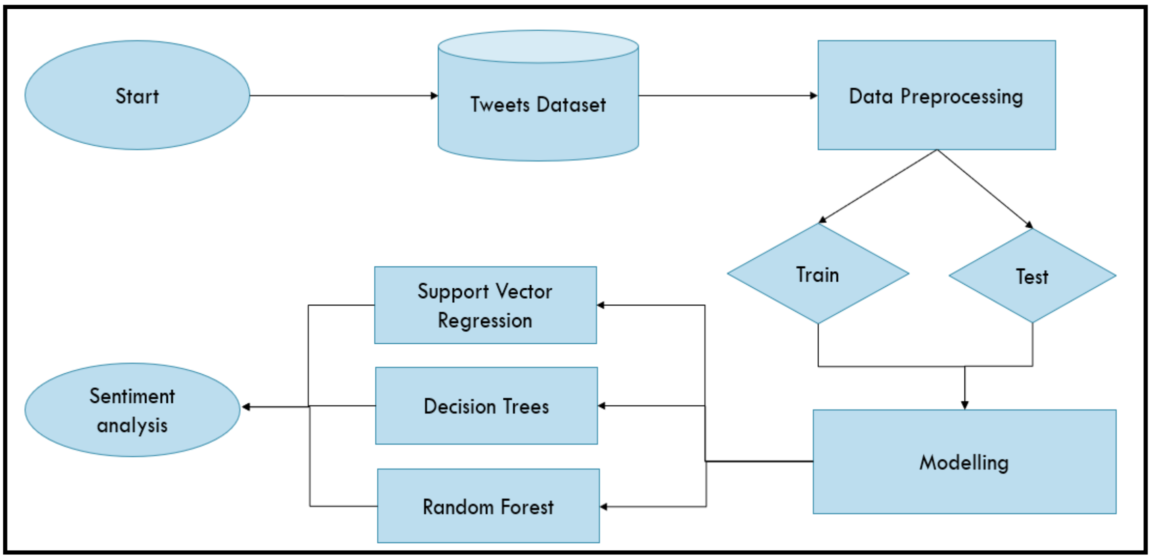


fig 1: Architecture of Proposed System

The proposed approach only analyses selected tweets from different domains using sentiment analysis. It is also strictly domain-restricted. It categorizes tweets as positve or negative or neutral using sentiment analysis. Below are the steps performed to analyze the data.

3.2 Data Cleaning

The format of the data that was gathered was raw data. The best method to ensure that information is accurate, foreseeable, and usable is to clean it up. Most of the time, datasets need to be cleaned up since they contain many noisy or undesirable data, often known as outliers. The presence of such outliers could produce unintended outcomes. Data cleaning ensures removing and modifying such data, creating a considerably more trustworthy and stable dataset. Data cleanup is possible in the following ways:

- Monitor errors - Error entry points or sources need to be continuously tracked and watched over. This will assist in cleaning up the damaged data.
- Accuracy validation - The data should be confirmed after cleaning the database. It is critical to research and uses various data technologies to assist in cleaning datasets.
- Avoid data duplication - It is necessary to identify duplicate data. Several AI techniques aid in the detection of recurrence in huge data sets.

There are numerous methods for cleaning datasets, some of which are discussed above. By utilizing these techniques, good, useable, and trustworthy datasets will be produced.

3.3 Data Pre-processing

Data preprocessing in Twitter sentiment analysis involves several steps to clean and prepare the textual data for accurate sentiment classification. Initially, the data undergoes several essential cleaning processes. To ensure uniformity in the text, lowercase conversion is applied across the board. This step avoids redundancy in classification due to variations in letter case. Noise removal is a crucial step that involves the elimination of special characters, punctuation, and symbols that don't contribute to the sentiment expressed. This step enhances the focus on relevant terms.

The preprocessing will consist of the following steps:

- Converting tweets to lowercase.
- Use spaces to replace at least two dots.
- Replace several spaces with a single space.
- Remove spaces and quotations at the end of tweets.

Stop words:

Stop words are terms that have no significance in search queries. For example, "I enjoy writing." After deleting the stop words, it becomes "like write." Stop words are "I" and "to." They should be avoided since they don't add much to the informative value of a sentence.

Tokenization:

Tokenization is a method for dividing a written document into tokens, relatively small units of text. A word, a word's fragment, or simple characters like punctuation can be a token. When dealing with text data, tokenization is crucial.

3.4 Proposed Algorithm:

3.4.1 Read NLTK Tweets:

Using this module we will read tweets from NLTK and then clean tweets by removing special symbols, stop words and then perform stemming (stemming means removing ing or tion from words for example ORGANIZATION word will become ORGANIZE after applying stem) on each words. Then we will calculate TFIDF vector.

3.4.2 Run Decision Tree 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.

Similarly we will build model for Random Forest and SVR to calculate their accuracy.

On internet opinion (sentiments on topic) mining is helping users in knowing the quality of any organization or products, if any user has good experience on any product or company then he will express good reviews/opinion and by seeing this opinion others users can know the quality of the product, in today's online social networks like twitter all peoples expressing their opinions and social networking sites developing new techniques to detect sentiments from this opinions, all existing techniques used to discover either Positive or Negative or Neutral sentiments from topics but this paper proposes 5 levels of sentiments detection such as High Positive, Moderate Positive, Neutral, High Negative and Moderate Negative. To detect sentiments author is using 4 Ordinal Regression machine learning algorithms such as Softmax, Decision Tree, Random Forest and Support Vector Regression.

Ordinal Regression means classifier used many independent variables to predict class of given data, In this paper also we give tweets as input and classifier predict sentiment by using all independent words from this tweets. Ordinal regression is a statistical technique that is used to predict behaviour of ordinal level dependent variables with a set of independent variables. The dependent variable is the order response category variable and the independent variable may be categorical or continuous.

Entropy:

- Entropy is a measure of impurity or disorder in a set of data.
- For a binary classification problem (e.g., positive and negative classes), the entropy $H(S)$ of a set S is calculated as follows:

$H(S)=-p{+}\log_{2}(p_{+}) - p\_\log_{2}(p\_)H(S)$

where:

- P_{+} is the proportion of positive instances in set S.
- P-is the proportion of negative instances in set S.
- $\log_{2}$ is the base-2 logarithm.

Information Gain:

- Information Gain measures the reduction in entropy after a dataset is split based on a particular attribute.
- For a dataset S with attribute A and values $\{v_1,v_2,\ldots,v_n\}$, the Information Gain $IG(S,A)$ is calculated as:

$IG(S,A)=H(S)-\sum i=1n|S||S_i|\cdot H(S_i)$

where:

- $|S|$ is the total number of instances in set S.
- S_i is the subset of instances in S where attribute A has value v_i.
- $|S_i|$ is the number of instances in S_i.
- $H(S_i)$ is the entropy of subset S_i.

From above algorithms Decision Tree is giving better prediction result and to train all algorithms we are using publicly available twitter dataset from NLTK library. We are using many features from NLTK (Natural Language Processing Tool Kit) library such as cleaning tweet text by removing special symbols, removing stop words (such as the, then, where etc.), word stemming which means removing ing, tionetc from words. After cleaning tweets then we will convert all tweets to BOG (Bag of Words Dictionary) and then convert BOG to vector by calculating TF/IDF (Term Frequency/Inverse Document Frequency).

3.4.3 Detect Sentiment Type:

Using this module we will upload test tweets and then application will apply train model on those test tweets to predict sentiment of that tweet.

- First, the text data needs to be represented in a numerical format. This is often done using techniques like Bag-of-Words, TF-IDF (Term Frequency-Inverse Document Frequency), or word embeddings.
- A sentiment classifier, such as a machine learning model, is trained using a labeled dataset. Each text sample in the training dataset is associated with a sentiment label (positive, negative, or neutral).
- The text is transformed into a feature vector X that represents the relevant characteristics of the text.
- The sentiment classifier is essentially a function *f* that takes the feature vector *X* as input and outputs a sentiment label.

  $f(X) \rightarrow \text{Sentiment Label}$
- In many cases, classifiers provide probability scores for each class. For example, in a binary sentiment analysis (positive/negative), the classifier might output probabilities $P(\text{Positive})$ and $P(\text{Negative})$.
- A decision rule is applied to determine the final sentiment label. For instance, if $P(\text{Positive}) > P(\text{Negative})$, the text might be classified as positive.

The simplified mathematical representation:

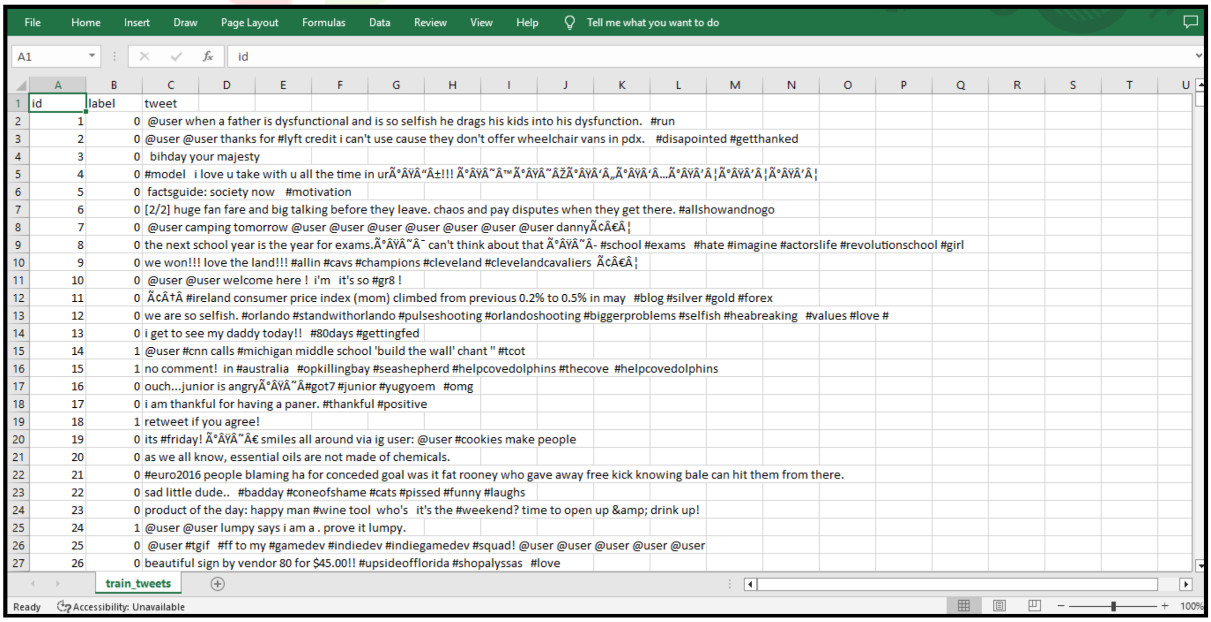
$f(X)=\text{argmax}_{\text{class}} P(\text{class}|X)$

Where:

- $f(X)$ is the sentiment classifier function.
- X is the feature vector representing the text.
- $\text{argmax}_{\text{class}}$ finds the class with the maximum probability.
- The probability $P(\text{class}|X)$ is often computed using techniques like logistic regression, Naive Bayes, or other machine learning algorithms.

3.5 Dataset:

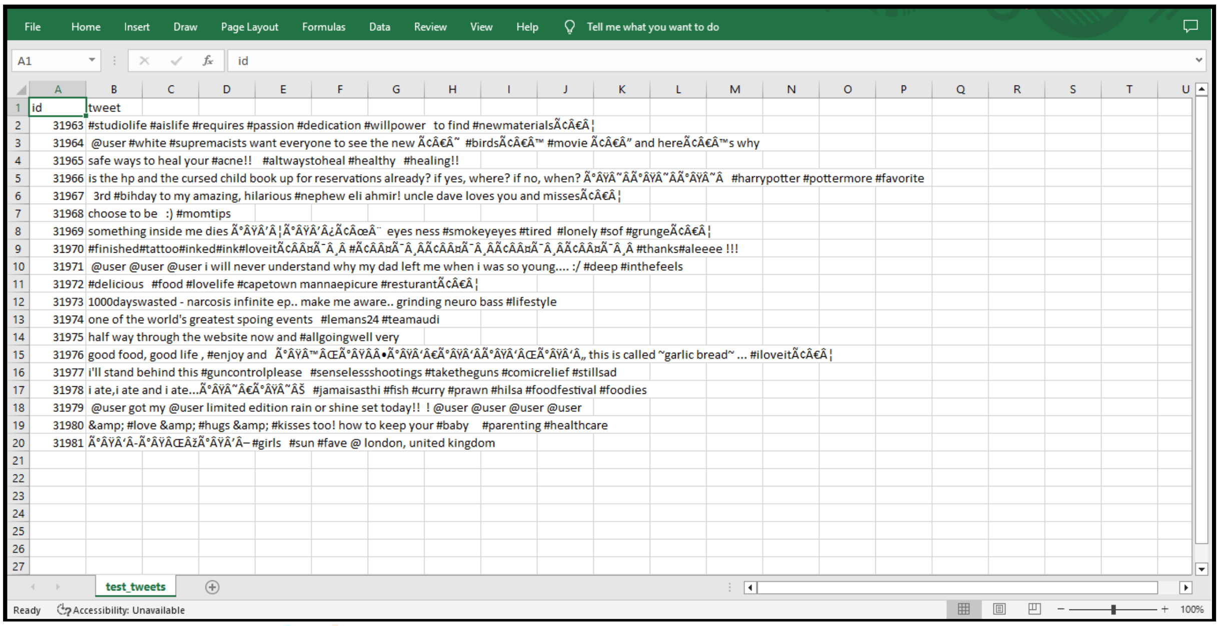
Training dataset:



id	label	tweet
1	0	@user when a father is dysfunctional and is so selfish he drags his kids into his dysfunction. #run
2	0	@user @user thanks for #lyft credit i can't use cause they don't offer wheelchair vans in pdx. #disapointed #getthanked
3	0	bihday your majesty
4	0	#model i love u take with u all the time in urÃ°ÂŸÂ“Â±!!! Ã°ÂŸÂ˜Â™Ã°ÂŸÂ˜ÂŽÃ°ÂŸÂ‘Â„Ã°ÂŸÂ‘Â…Ã°ÂŸÂ’Â¦Ã°ÂŸÂ’Â¦Ã°ÂŸÂ’Â¦
5	0	factsguide: society now #motivation
6	0	[2/2] huge fan fare and big talking before they leave. chaos and pay disputes when they get there. #allshowandnogo
7	0	@user camping tomorrow @user @user @user @user @user @user @user dannyÃ¢Â€Â¦
8	0	the next school year is the year for exams.Ã°ÂŸÂ˜Â¯ can't think about that Ã°ÂŸÂ˜Â #school #exams #hate #imagine #actorslife #revolutionschool #girl
9	0	we won!!! love the land!!! #allin #cavs #champions #cleveland #clevelandcavaliers Ã¢Â€Â¦
10	0	@user @user welcome here ! i'm it's so #gr8 !
11	0	Ã¢Â†Â #ireland consumer price index (mom) climbed from previous 0.2% to 0.5% in may #blog #silver #gold #forex
12	0	we are so selfish. #orlando #standwithorlando #pulseshooting #orlandoshooting #biggerproblems #selfish #heabreaking #values #love #
13	0	i get to see my daddy today!! #80days #gettingfed
14	1	@user #cnn calls #michigan middle school 'build the wall' chant " #tcot
15	1	no comment! in #australia #opkillingbay #seashepherd #helpcovedolphins #thecove #helpcovedolphins
16	0	ouch...junior is angryÃ°ÂŸÂ˜Â#got7 #junior #yugyoem #omg
17	0	i am thankful for having a paner. #thankful #positive
18	1	retweet if you agree!
19	0	its #friday! Ã°ÂŸÂ˜Â€ smiles all around via ig user: @user #cookies make people
20	0	as we all know, essential oils are not made of chemicals.
21	0	#euro2016 people blaming ha for conceded goal was it fat rooney who gave away free kick knowing bale can hit them from there.
22	0	sad little dude.. #badday #coneofshame #cats #pissed #funny #laughs
23	0	product of the day: happy man #wine tool who's it's the #weekend? time to open up &amp; drink up!
24	1	@user @user lumpy says i am a . prove it lumpy.
25	0	@user #tgif #ff to my #gamedev #indiedev #indiegamedev #squad! @user @user @user @user @user
26	0	beautiful sign by vendor 80 for $45.00!! #upsideofflorida #shopalyssas #love

fig 2: training dataset

Testing dataset:



id	tweet
31963	#studiolife #aislife #requires #passion #dedication #willpower to find #newmaterialsÃ¢Â€Â¦
31964	@user #white #supremacists want everyone to see the new Ã¢Â€Â˜ #birdsÃ¢Â€Â™ #movie Ã¢Â€Â” and hereÃ¢Â€Â™s why
31965	safe ways to heal your #acne!! #altwaystoheal #healthy #healing!!
31966	is the hp and the cursed child book up for reservations already? if yes, where? if no, when? Ã°ÂŸÂ˜ÂÃ°ÂŸÂ˜ÂÃ°ÂŸÂ˜Â #harrypotter #pottermore #favorite
31967	3rd #bihday to my amazing, hilarious #nephew eli ahmir! uncle dave loves you and missesÃ¢Â€Â¦
31968	choose to be :) #momtips
31969	something inside me dies Ã°ÂŸÂ’Â¦Ã°ÂŸÂ’Â¿Ã¢ÂœÂ¨ eyes ness #smokeyeyes #tired #lonely #sof #grungeÃ¢Â€Â¦
31970	#finished#tattoo#inked#ink#loveitÃ¢ÂÂ¤Ã¯Â¸Â #Ã¢ÂÂ¤Ã¯Â¸ÂÃ¢ÂÂ¤Ã¯Â¸ÂÃ¢ÂÂ¤Ã¯Â¸ÂÃ¢ÂÂ¤Ã¯Â¸Â #thanks#aleeee !!!
31971	@user @user @user i will never understand why my dad left me when i was so young.... :/ #deep #inthefeels
31972	#delicious #food #lovelife #capetown mannaepicure #resturantÃ¢Â€Â¦
31973	1000dayswasted - narcosis infinite ep.. make me aware.. grinding neuro bass #lifestyle
31974	one of the world's greatest spoing events #lemans24 #teamaudi
31975	half way through the website now and #allgoingwell very
31976	good food, good life , #enjoy and Ã°ÂŸÂ™ÂŒÃ°ÂŸÂÂ•Ã°ÂŸÂ‘Â€Ã°ÂŸÂ‘ÂÃ°ÂŸÂ‘ÂŒÃ°ÂŸÂ‘Â„ this is called ~garlic bread~ ... #iloveitÃ¢Â€Â¦
31977	i'll stand behind this #guncontrolplease #senselessshootings #taketheguns #comicrelief #stillsad
31978	i ate,i ate and i ate...Ã°ÂŸÂ˜Â€Ã°ÂŸÂ˜ÂŠ #jamaisasthi #fish #curry #prawn #hilsa #foodfestival #foodies
31979	@user got my @user limited edition rain or shine set today!! ! @user @user @user @user
31980	&amp; #love &amp; #hugs &amp; #kisses too! how to keep your #baby #parenting #healthcare
31981	Ã°ÂŸÂ‘Â-Ã°ÂŸÂŒÂžÃ°ÂŸÂ’Â– #girls #sun #fave @ london, united kingdom

fig 3: testing dataset

IV. RESULTS:

The code begins by importing necessary libraries, including Tkinter for GUI, NLTK for natural language processing, scikit-learn for machine learning, and TextBlob for text processing.

The Python code implements a sentiment analysis tool with a Tkinter-based graphical user interface. It utilizes natural language processing techniques and machine learning models to classify tweets into positive, negative, or neutral sentiments. The tool first loads tweets from the NLTK Twitter dataset, processes them, and splits them into training and test sets. Three machine learning models—Support Vector Machine (SVM), Random Forest, and Decision Tree—are trained on the dataset. The accuracy of each model is displayed in the GUI. Users can then upload a file containing tweets for sentiment analysis, and the tool classifies each tweet, providing results such as "High Positive," "Moderate Positive," "Neutral," "Moderate Negative," or "High Negative" based on probability values. Some adjustments, particularly in the Random Forest accuracy calculation, may be needed for optimal performance.

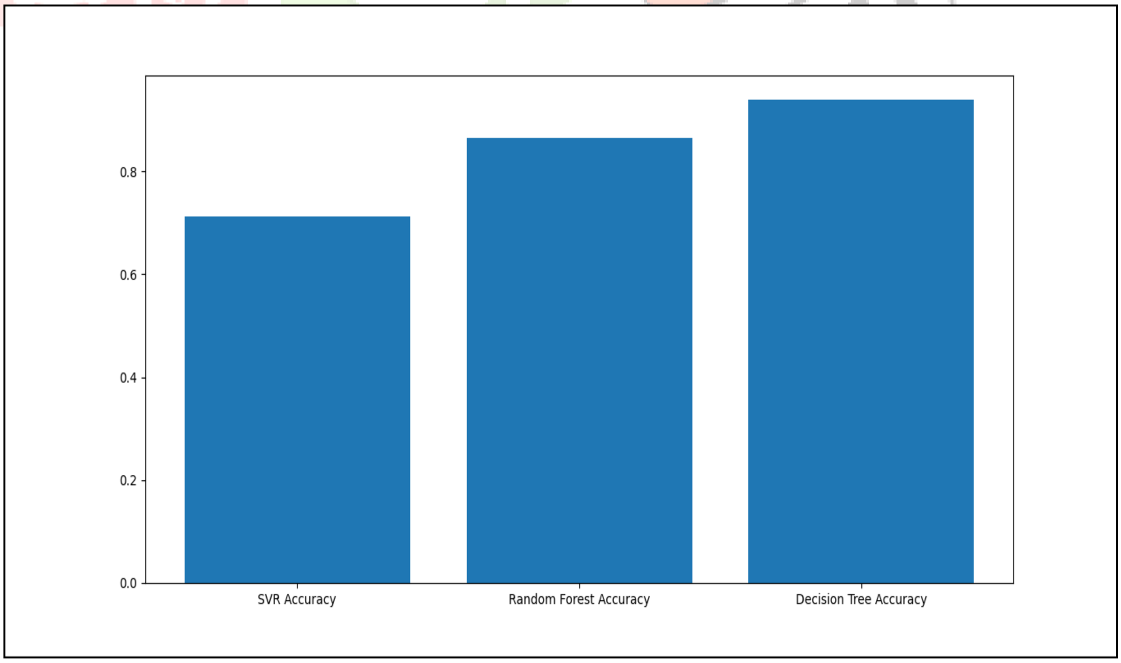
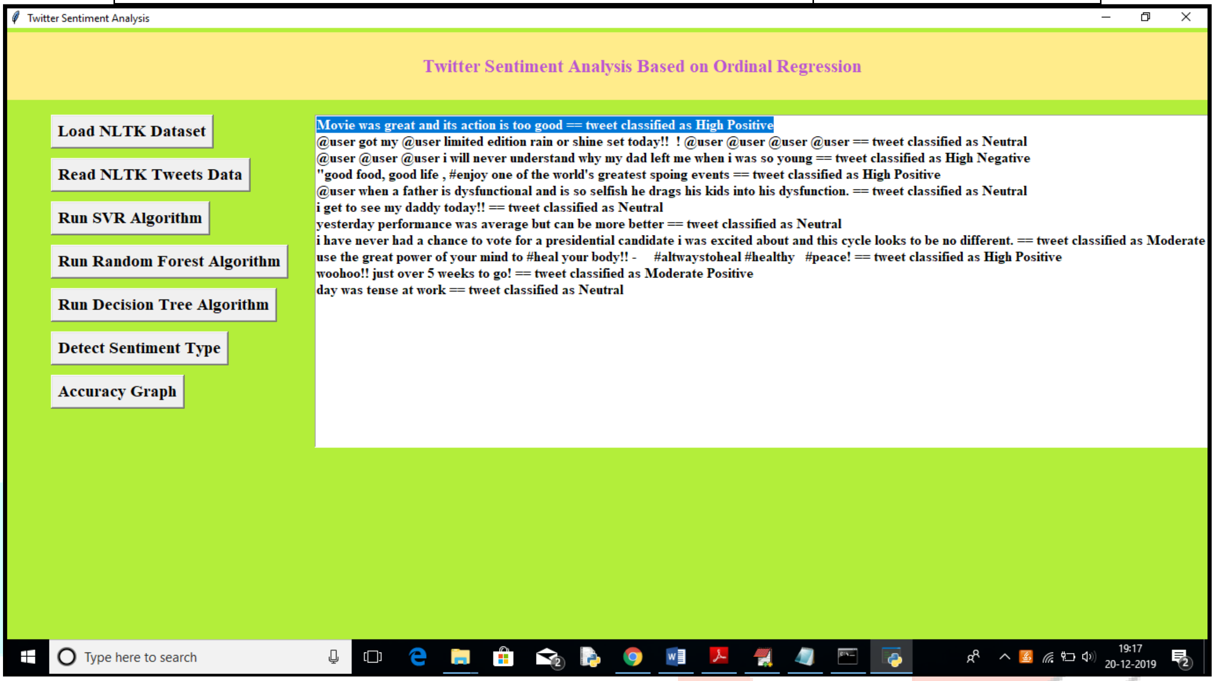


fig 4: accuracy graph

Table 2: Sentiment Type

Algorithm	**Accuracy**
Support Vector Regression	71.3%
Random Forest	88.5%
Decision Tree	93.5%


Twitter Sentiment Analysis
Twitter Sentiment Analysis Based on Ordinal Regression
Load NLTK Dataset
Read NLTK Tweets Data
Run SVR Algorithm
Run Random Forest Algorithm
Run Decision Tree Algorithm
Detect Sentiment Type
Accuracy Graph
Type here to search
19:17
20-12-2019

```
Movie was great and its action is too good == tweet classified as High Positive
@user got my @user limited edition rain or shine set today!!  ! @user @user @user @user == tweet classified as Neutral
@user @user @user i will never understand why my dad left me when i was so young == tweet classified as High Negative
"good food, good life , #enjoy one of the world's greatest spoing events == tweet classified as High Positive
@user when a father is dysfunctional and is so selfish he drags his kids into his dysfunction. == tweet classified as Neutral
i get to see my daddy today!! == tweet classified as Neutral
yesterday performance was average but can be more better == tweet classified as Neutral
i have never had a chance to vote for a presidential candidate i was excited about and this cycle looks to be no different. == tweet classified as Moderate
use the great power of your mind to #heal your body!! -     #altwaystoheal #healthy   #peace! == tweet classified as High Positive
woohoo!! just over 5 weeks to go! == tweet classified as Moderate Positive
day was tense at work == tweet classified as Neutral
```

Fig 5: Sentiment Type

The above picture shows the website of a Twitter Sentiment Analysis tool. The tool uses three machine learning algorithms to classify tweets into four sentiment categories: High Positive, Moderate Positive, Neutral, and High Negative. The algorithms are:

- Support Vector Regression
- Random Forest
- Decision Tree

Load NLTK Dataset

1. "Movie was great and its action is too good" - High Positive
2. "@user got my @user limited edition rain or shine set today!!! @user user useruser" - Neutral
3. "@user useruseri will never understand why my dad left me when i was so young" - High Negative
4. "@user when a father is dysfunctional and is so selfish he drags his kids into his dysfunction." - Neutral

Overall, it shows a Twitter Sentiment Analysis tool that uses machine learning to classify tweets into four sentiment categories. The tool has an accuracy of 93% using the decision algorithm

V. CONCLUSION:

This study undertakes the ambitious task of elucidating the sentiment analysis of Twitter data through the lens of ordinal regression, employing a variety of machine learning techniques. The proposed approach centers on constructing a robust balancing and scoring model, followed by the classification of tweets into distinct ordinal classes using machine learning classifiers. The classifiers employed in this investigation include Multinomial Logistic Regression, Support Vector Regression, Decision Trees, and Random Forest. The foundation for this research is laid upon an optimized Twitter dataset, publicly available within the NLTK corpora resources.The experimental findings reveal compelling insights into the performance of the machine learning classifiers. Notably, Support Vector Regression and Random Forest exhibit comparable accuracies, surpassing that of the Multinomial Logistic Regression classifier. Intriguingly, the Decision Tree classifier emerges as the frontrunner with the highest accuracy recorded at an impressive 91.81%. These results underscore the efficacy of the proposed model in detecting ordinal regression within the realm of Twitter sentiment analysis, showcasing a commendable level of accuracy. To comprehensively evaluate the model's performance, multiple metrics have been employed, including accuracy, Mean Absolute Error, and Mean Squared Error. These metrics collectively contribute to a holistic understanding of the model's precision and reliability in discerning ordinal sentiment classes within the dynamic and nuanced context of Twitter data.As a forward-looking trajectory, the study outlines strategic plans for refining and expanding the proposed approach. The incorporation of bigrams and trigrams is slated as an avenue for improvement, aiming to capture more nuanced linguistic patterns inherent in Twitter data. Additionally, the research endeavors to explore a broader spectrum of machine learning techniques, delving into the realms of deep learning methodologies. Future investigations are anticipated to explore the application of Deep Neural Networks, Convolutional Neural Networks, and Recurrent Neural Networks, seeking to unlock further potential in enhancing the model's predictive capabilities.

In summary, the study presents a comprehensive and insightful exploration of ordinal regression in Twitter sentiment analysis, demonstrating the viability of the proposed model. The amalgamation of balancing and scoring, coupled with the utilization of diverse machine learning classifiers, positions this approach as a robust contender in the field. The documented results provide a strong foundation for future enhancements and expansions, setting the stage for a deeper understanding of sentiment dynamics within the Twitter landscape.

VI. REFERENCES

[1] Palomino, M.A.; Aider, F. Evaluating the Effectiveness of Text Pre-Processing in Sentiment Analysis. Appl. Sci. 2022, 12, 8765. https://doi.org/10.3390/app12178765

[2] Hassan, Saeed-Ul, Saleem, Aneela, Soroya, Saira Hanif, Safder, Iqra, Iqbal, Sehrish, Jamil, Saqib, Bukhari, Faisal, Aljohani, Naif Radi and Nawaz, Raheel (2021) Sentiment analysis of tweets through Altmetrics: a machine learning approach. Journal of Information Science, 47 (6). pp. 712-726. ISSN 0165-5515

[3] Sumith Pevekar , Aayush Pandey , Naresh Alwala , Prakash Parmar, 2021, A Machine Learning based Approach for Determining Consumer Purchase Intention using Tweets, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 10, Issue 09 (September 2021),DOI : 10.17577/IJERTV10IS090244

[4] S. Chaurasia, S. Sherekar and V. Thakare, "Twitter Sentiment Analysis using Natural Language Processing," 2021 International Conference on Computational Intelligence and Computing Applications (ICCICA), Nagpur, India, 2021, pp. 1-5, doi: 10.1109/ICCICA52458.2021.9697136.

[5] A. Roy and M. Ojha, "Twitter sentiment analysis using deep learning models," 2020 IEEE 17th India Council International Conference (INDICON), New Delhi, India, 2020, pp. 1-6, doi: 10.1109/INDICON49873.2020.9342279.

[6] L. Mandloi and R. Patel, "Twitter Sentiments Analysis Using Machine Learninig Methods," 2020 International Conference for Emerging Technology (INCET), Belgaum, India, 2020, pp. 1-5, doi: 10.1109/INCET49848.2020.9154183.

[7] C. Kariya and P. Khodke, "Twitter Sentiment Analysis," 2020 International Conference for Emerging Technology (INCET), Belgaum, India, 2020, pp. 1-3, doi: 10.1109/INCET49848.2020.9154143.

[8] N. F. Alshammari and A. A. AlMansour, "State-of-the-art review on Twitter Sentiment Analysis," 2019 2nd International Conference on Computer Applications & Information Security (ICCAIS), Riyadh, Saudi Arabia, 2019, pp. 1-8, doi: 10.1109/CAIS.2019.8769465.

[9] S. A. El Rahman, F. A. AlOtaibi and W. A. AlShehri, "Sentiment Analysis of Twitter Data," 2019 International Conference on Computer and Information Sciences (ICCIS), Sakaka, Saudi Arabia, 2019, pp. 1-4, doi: 10.1109/ICCISci.2019.8716464.

[10] A. P. Rodrigues, A. Rao and N. N. Chiplunkar, "Sentiment Analysis of Real Time Twitter Data Using Big Data Approach," 2017 2nd International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS), Bengaluru, India, 2017, pp. 1-6, doi: 10.1109/CSITSS.2017.8447656.

[11] Swathi, A. et al. (2023). A Reliable Novel Approach of Bio-Image Processing—Age and Gender Prediction. In: Reddy, K.A., Devi, B.R., George, B., Raju, K.S., Sellathurai, M. (eds) Proceedings of Fourth International Conference on Computer and Communication Technologies. Lecture Notes in Networks and Systems, vol 606. Springer, Singapore. https://doi.org/10.1007/978-981-19-8563-8_31

[12] Swathi, A., Ashwani Kumar, V. Swathi, Y. Sirisha, D. Bhavana, Shaik Abdul Latheef, A. Abhilash, and G. Mounika. "Driver Drowsiness Monitoring System Using Visual Behavior And Machine Learning." In *2022 5th International Conference on Multimedia, Signal Processing and Communication Technologies (IMPACT)*, pp. 1-4. IEEE, 2022.

[13] Gowroju, Swathi, and Sandeep Kumar. "Robust deep learning technique: U-net architecture for pupil segmentation." In *2020 11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pp. 0609-0613. IEEE, 2020.

[14] Gowroju, Swathi, and Sandeep Kumar. "Robust pupil segmentation using UNET and morphological image processing." In *2021 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC)*, pp. 105-109. IEEE, 2021.

[15] Gowroju, Swathi, and Sandeep Kumar. "Robust deep learning technique: U-net architecture for pupil segmentation." In *2020 11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pp. 0609-0613. IEEE, 2020.

[16] Gowroju, Swathi, K. Sravani, N. Santhosh Ramchandar, D. Sai Kamesh, and J. Nasrasimha Murthy. "Robust Indian Currency Recognition Using Deep Learning." In *Advanced Informatics for Computing Research: 4th International Conference, ICAICR 2020, Gurugram, India, December 26–27, 2020, Revised Selected Papers, Part I 4*, pp. 477-486. Springer Singapore, 2021.

[17] Swathi, A., and Shilpa Rani. "Intelligent fatigue detection by using ACS and by avoiding false alarms of fatigue detection." In *Innovations in Computer Science and Engineering: Proceedings of the Sixth ICICSE 2018*, pp. 225-233. Springer Singapore, 2019.