ASPECT BASED SENTIMENT ANALYSIS FOR E-COMMERCE REVIEWS USING MACHINE LEARNING

A Project Report
In the partial fulfilment of the award of the degree of

Bachelor Of Technology

Under

Ardent Computech Pvt. Ltd.

Submitted by:
SANJUKTA CHAKRABORTY
TANMAYEE GOUDA
ANANYA KARMAKAR
SAYAN DAS

FUTURE INSTITUTE OF ENGINEERING

AND MANAGEMENT

SONARPUR STATION ROAD, KOLKATA

700150

CERTIFICATE FROM THE MENTOR

This is to certify that **SANJUKTA CHAKRABORTY, TANMAYEE GOUDA, ANANYA KARMAKAR AND SAYAN DAS** has completed the project titled "ASPECT BASED SENTIMENT ANALYSIS FOR E-COMMERCE REVIEWS USING MACHINE LEARNING" under my supervision during the period from December to March which is in partial fulfilment of requirements for the award of the Bachelor of Computer Science and Engineering through Ardent Computech Pvt. Ltd.

DATE: **Signature of the Mentor.**

ACKNOWLEDGMENT

I take this opportunity to express my deep gratitude and sincerest thanks to my project mentor, **Mr.Joyjit Guha Biswas** for giving the most valuable suggestions, helpful guidance, and encouragement in the execution of this project work.

I would like to give a special mention to my colleagues. Last but not least I am grateful to all the faculty members of the **Ardent Computech Pvt.Ltd.** for their support.

(Note: All entries of the proforma of approval should be filled up with appropriate and complete information of approval in any respect will be summarily rejected.)

1. Name of the Student With Group:

   i. SANJUKTA CHAKRABORTY
   ii. TANMAYEE GOUDA
   iii. ANANYA KARMAKAR
   iv. SAYAN DAS

2. Title of the Project : **ASPECT BASED SENTIMENT ANALYSIS FOR E-COMMMERCE REVIEWS USING MACHINE LEARNING**

3. Name and Address of the Guide : **MR. JOYJIT GUHA BISWAS**

   Sr. Subject Matter Expert & Technical Head (Python) Ardent Computech Pvt. Ltd. (An ISO 9001:2015 Certified) Module-132, SDF Building, Sector-V, Kolkata-700091

4. Educational Qualification of the Guide : Ph.d* ☐ M.tech* ☐ B.E*/B.Tech * ☐ MCA* ☐ M.Sc* ☐

5. Working and Teaching experience of the Guide ..... **Years**

6. Software used in the Project:

   a. **Google collab**
   b. **Python**

Signature of the Guide

Date:

Name: Mr. Joyjit Guha Biswas

Subject Matter Expert

Signature, Designation, Stamp of the Project Proposal Evaluator

APPROVED ☐ NOT APPROVED ☐

PROJECT RESPONSIBILITY FORM

Aspect Based Sentiment Analysis For E-Commerce Reviews

SERIAL NO.	NAME	RESPONSIBILITY
1	SANJUKTA CHAKRABORTY	DATA COLLECTION AND PREPROCESSING
2	TANMAYEE GOUDA	VECTORIZATION AND MODEL CREATION
3	ANANYA KARMAKAR	MODEL TESTING AND CALCULATING ACCURACY
4	SAYAN DAS	DOCUMENTATION

SELF- CERTIFICATE

This is to certify that the dissertation/project proposal entitled "Aspect Based Sentiment Analysis For E-Commerce Reviews Using Machine Learning" is done by us, is an Information Technology under the guidance of Mr. Joyjit Guha Biswas. The matter embodied in this project work has not been submitted earlier for award of any certificate to the best of our knowledge and belief.

Name of the Students:

1. SANJUKTA CHAKRABORTY
2. TANMAYEE GOUDA
3. ANANYA KARMAKAR
4. SAYAN DAS

Signature of the students:

1. Sanjukta Chakraborty
2. Tanmayee Gouda
3. Ananya Karmakar
4. Sayan Das

CERTIFICATE BY GUIDE

This is to certify that this project entitled "Aspect Based Sentiment Analysis For E-Commerce Reviews Using Machine Learning" submitted in partial fulfillment of the certificate of Bachelor of Computer Science and Engineering through Ardent Computech Pvt. Ltd., done by the

Group Members:

1. Sanjukta Chakraborty
2. Tanmayee Gouda
3. Ananya Karmakar
4. Sayan Das

is an authentic work carried out under my guidance & best of our knowledge and belief.

1. Sanjukta Chakraborty
2. Tanmayee Gouda
3. Ananya Karmakar
4. Sayan Das

Signature of the students
Date:

Signature of the Guide
Date:

CERTIFICATE OF APPROVAL

This is to certify that this proposal of Minor project, entitled "**Aspect Based Sentiment Analysis For E-Commerce Reviews Using Machine Learning**" is a record of bona-fide work, carried out by:

1. Sanjukta Chakraborty, 2. Tanmayee Gouda , 3. Ananya Karmakar, 4. Sayan Das under my supervision and guidance through the Ardent Computech Pvt. Ltd. In my opinion, the report in its present form is in partial fulfillment of all the requirements, as specified by the Future Institute of Engineering and Management Collage (ECE Department) as per regulations of the Ardent Computech Pvt. Ltd. In fact, it has attained the standard, necessary for submission. To the best of my knowledge, the results embodied in this report, are original in nature and worthy of incorporation in the present version of the report for Bachelor of Technology.

Guide/Supervisor

Mr. Joyjit Guha Biswas

Subject Matter Expert & Technical Head (Python)
Ardent Computech Pvt. Ltd. (An ISO 9001:2015 Certified) Module-132, SDF Building
Salt Lake Sector-V, Kolkata - 700 091

External Examiner(s)

Head of ECE Department
(Future Institute of Engineering and Management Collage)

ASPECT BASED SENTIMENT ANALYSIS FOR E-COMMERCE REVIEWS

TABLE OF CONTENT

1.1. PROJECT OVERVIEW

Aspect Based Sentiment Analysis For E-Commerce Reviews Using Machine Learning is a critical task in medical image analysis. Early and accurate detection is crucial for effective treatment planning. This project aims to develop a machine-learning-based model to automatically detect sentiment from text comments.

1.1.1. PROBLEM STATEMENT

The manual detection of sentiment analysis is time-consuming, prone to human error, and often requires expertise. There is a need for an automated system to assist people and social media and shopping apps in accurately and efficiently detecting sentiments.

1.1.2. OBJECTIVE

The primary objective of this project is to develop a robust machine learning model capable of accurately detecting ABSA correctly for a text sentence/comment. The model should be able to classify texts into different sentiment types.

SCOPE

This project focuses on developing and evaluating a machine learning model for sentiment analysis using a specific dataset. The scope includes:

- Data preprocessing and augmentation
- Model architecture selection and training
- Model evaluation using relevant metrics
- Exploratory analysis of model performance

1.2. DATASET DESCRIPTION

1.2.1. SOURCE OF DATASET:

Kaggle :- https://www.kaggle.com/code/aaroha33/e-commerce-sentiment-analysis

1.2.2. DATA PREPROCESSING STEPS

Data preprocessing is a crucial step in ensuring the quality and consistency of the dataset, which directly impacts the model's performance. The following preprocessing steps were undertaken:

IMPORT ALL DEPENDENCIES:

Import All Dependencies here

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import re
import nltk
nltk.download('book')
from nltk.book import *
from scipy.special import softmax
from sklearn.model_selection import train_test_split
from nltk.stem.porter import PorterStemmer
from nltk.corpus import stopwords
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import classification_report, confusion_matrix
import string
from textblob import TextBlob
from sklearn import preprocessing
import warnings
warnings.filterwarnings('ignore')
```

DATASET LOADING:

Reading the Dataset

+ Code + Text

```
[ ] df= pd.read_csv('Ardent_Project.csv')
    df
```

	id	name	asins	brand	categories	keys	manufacturer	reviews.date	reviews.dateAdded	revie
0	AVqkIhwDv8e3D1O-lebb	All-New Fire HD 8 Tablet, 8 HD Display, Wi-Fi,...	B01AHB9CN2	Amazon	Electronics,iPad & Tablets,All Tablets,Fire Ta...	841667104676,amazon/53004484,amazon/b01ahb9cn2...	Amazon	2017-01-13T00:00:00.000Z	2017-07-03T23:33:15Z	07T09:04:0 04-30T0
1	AVqkIhwDv8e3D1O-lebb	All-New Fire HD 8 Tablet, 8 HD Display, Wi-Fi,...	B01AHB9CN2	Amazon	Electronics,iPad & Tablets,All Tablets,Fire Ta...	841667104676,amazon/53004484,amazon/b01ahb9cn2...	Amazon	2017-01-13T00:00:00.000Z	2017-07-03T23:33:15Z	07T09:04:0 04-30T0
2	AVqkIhwDv8e3D1O-lebb	All-New Fire HD 8 Tablet, 8 HD	B01AHB9CN2	Amazon	Electronics,iPad & Tablets,All Tablets,Fire Ta...	841667104676,amazon/53004484,amazon/b01ahb9cn2...	Amazon	2017-01-13T00:00:00.000Z	2017-07-03T23:33:15Z	07T09:04:0 04-30T0

- Data is loaded from the dataset in their original format.

DATASET CLEANING AND CHECKING FOR NULL VALUES:

- The whole dataframe is checked for null values using .isnull().sum() function and if there then try to drop it or fill it according to our need.

```
df.isnull().sum()
```

	0
name	0
brand	0
categories	0
manufacturer	0
date	0
reviews.dateAdded	0
reviews.doRecommend	0
rating	0
Product links	0
reviews	0
title	0

dtype: int64

DESCRIBING THE DATA:

```
[22] df.describe()
```

	rating	length
count	34660.000000	34660.000000
mean	4.584882	159.043566
std	0.735424	185.836598
min	1.000000	3.000000
25%	4.000000	70.000000
50%	5.000000	106.000000
75%	5.000000	183.000000
max	5.000000	10670.000000

TOKEN GENERATION:

```
[ ] string.punctuation

'!"#$%&\'()*+,-./:;<=>?@[\\]^_`{|}~'

[ ] def remove_punctuation(rev_list):
        rev_list = ''.join([word for word in rev_list if word not in string.punctuation])
        return rev_list

    df['cleaned_reviews'] = df['cleaned_reviews'].apply(lambda x: remove_punctuation(x))
```

- In this step the text messages are splitted into tokens for easy classification for which four parts of nltk module are used:
    - punkt
    - wordnet

STEMMING:

```
corpus=[]
def clean_text(rev_list):
    rev_list= re.sub(r'<.*?>', '', rev_list)  # Remove HTML tags
    rev_list = re.sub(r'[^a-zA-Z0-9 ]', '', rev_list)  # Remove special characters
    rev_list = rev_list.lower()
    rev_list = rev_list.split()
    ps = PorterStemmer()
    rev_list = [ps.stem(word) for word in rev_list if not word in set(stopwords.words('english'))]
    rev_list = ' '.join(rev_list)
    corpus.append(rev_list)
    return rev_list

df['cleaned_reviews'] = df['reviews'].apply(clean_text)
```

- Here every sentiments are labelled into three categories for easy classification for the machine learning algorithm.
  - Positive
  - Negative
  - Neutral
- It will improve the overall accuracy of the model and leads to easy sentiment detection.

LABEL ENCODING:

```
positive= 28518/len(df)*100
negative= 4401/len(df)*100
neutral= 1741/len(df)*100

labels = ['Positive', 'Negative', 'Neutral']
sizes = [positive, negative, neutral]
plt.figure(figsize=(12,7))
plt.pie(sizes, labels=labels,autopct='%1.1f%%',shadow=True, startangle=140)
plt.show()
```

- Here all the sentiments are encoded into numerical values using Label Encoder to avoid any kind of string related error.

DATA SPLITTING:

```
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=2)

print(x.shape, x_train.shape, x_test.shape)
```

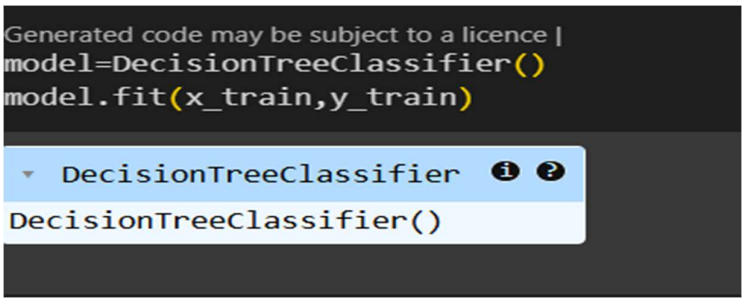
- The dataset was divided into training and testing sets with a ratio of [insert ratio, e.g., 80:20]. This split ensures adequate data for model training, evaluation, and final assessment.

2. METHODOLOGY

2.1 MODEL ARCHITECTURE

2.1.1 CHOSEN ARCHITECTURE:

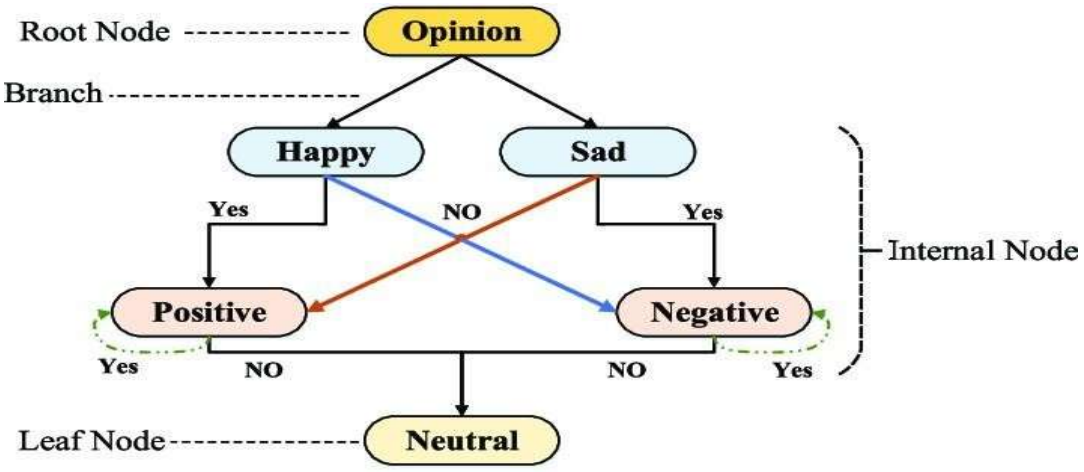
The chosen architecture for this project is Decision tree among Random Forest, Multinomial Naïve Bayes theorem, Logistic regression, LSTM techniques and Decision Tree as we have got more accuracy in case of Decision Tree.



ABOUT MODEL:

A decision tree is a flow-chart like diagram that shows how a series of decisions lead to different outcomes. It is a popular machine learning tool for classification and regression tasks.

- A decision tree starts with a root node, which has no incoming branches. ○ From the root node, the tree asks a series of yes/no questions, or conditions, to split the data into subsets.
- Each Branch represents a possible outcome of a decision.
- Each leaf node represents a classification or prediction.



BENEFIT OF THE MODEL:

- Decision trees are easy to interpret and can handle categorical features.
- They can capture non-linearities and feature interactions.
- Decision trees can help different groups in an organization understand why a decision was made.

2.1.2. MODEL COMPILATION

The model is compiled using the following parameters:

- **Confusion Metrics:** It is used to see that how many text sentiments are correctly and how many of that are wrongly classified using True Positive (TR), True Negative (TN), False Positive (FP) and False Negative (FN) values.
- **Classification Report:** It is used to show the **Accuracy**, **Precision, Recall** and **F1-Score** to make the all over summary of the model for the particular dataset.

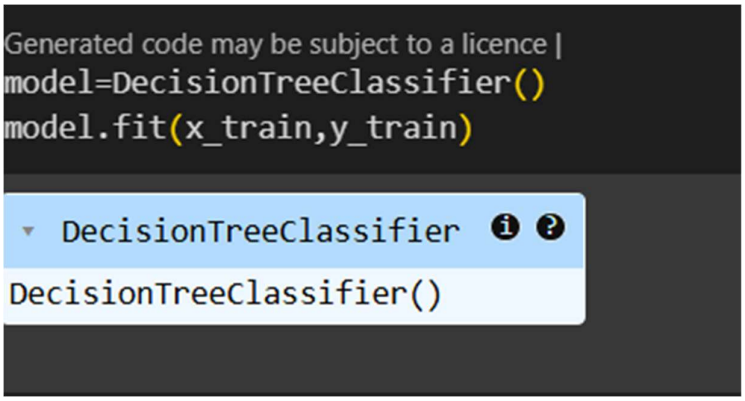
```
model = DecisionTreeClassifier()
model.fit(x_train_transformed, y_train)  # Train the model

x_test_prediction = model.predict(x_test_transformed)  # Predict
accuracy = accuracy_score(y_test, x_test_prediction)  # Calculate accuracy

print(f"Model Accuracy: {accuracy:.2f}")
```

```python
# F1 Score Calculation
f1 = f1_score(y_test, x_test_prediction, average="weighted")  # Weighted for multi-class

# Classification Report
report = classification_report(y_test, x_test_prediction)


print(f"F1 Score: {f1:.2f}")
print("\nClassification Report:\n", report)
```

```
F1 Score: 0.95

Classification Report:
              precision    recall  f1-score   support

           0       0.98      0.47      0.64       352
           1       0.91      0.92      0.91       868
           2       0.97      1.00      0.98      5712

    accuracy                           0.96      6932
   macro avg       0.95      0.80      0.84      6932
weighted avg       0.96      0.96      0.95      6932
```

2.2 TRAINING AND EVALUATION

2.2.1. TRAINING PROCESS:



The model was trained using the x_train and y_train dataframes.

The training process involved iteratively making the root nodes to see in which case the accuracy wil be highest and that case will be considered for further testing and evaluation of the model.

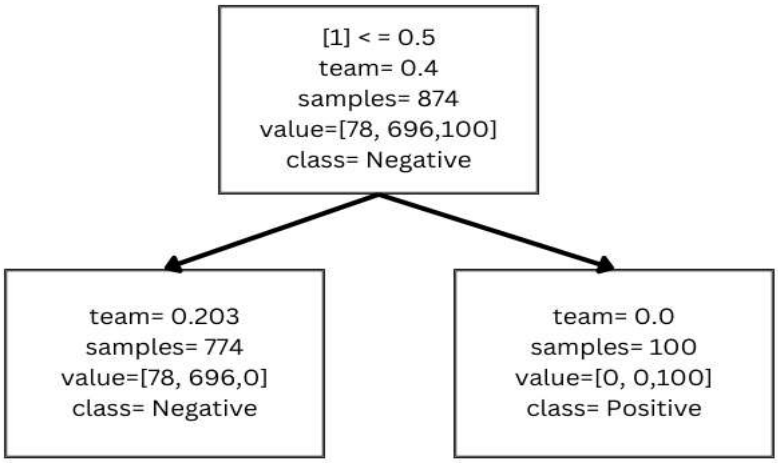
2.2.2 EVALUATION METRICS

To assess the model's performance, the following metrics were employed:

- **Accuracy:** The proportion of correctly classified instances.
- **Precision:** The proportion of positive predictions that were truly positive.
- **Recall:** The proportion of actual positives that were correctly identified as positive.
- **F1-Score:** The measure of that how many times a model predicts correctly among the entire dataset.

These metrics provide a comprehensive evaluation of the model's ability to correctly classify text sentiments.

3. RESULTS:



3.1. MODEL PERFORMANCE ANALYSISBASED ON THE PROVIDED CONFUSION METRICS AND CLASSIFICATION REPORT:

- From the Confusion matrix we can clearly see that the most of the values are in the TP and TN side on metrics means the prediction has a higher accuracy rate.
- From the classification report we can see that the f1-score for every category is pretty high. So we can tell that the model is working as a good one.

MODEL TESTING:

```
y_pred=model.predict(x_test_transformed)
y_pred
```

```
217/217 ━━━━━━━━━━━━━━━━━━━━ 692s 3s/step
array([[1.],
       [1.],
       [1.],
       ...,
       [1.],
       [1.],
       [1.]], dtype=float32)
```

```
accuracy = accuracy_score(y_test, x_test_prediction)  # Calculate accuracy

print(f"Model Accuracy: {accuracy:.2f}")
```

```
Model Accuracy: 0.97
```

Overview:

The provided decision tree plotting gives us some amount of insight about the dataset.

Key Observations:

- As we can see initially the team index is about 0.4 which is pretty low score and as we know a lower team index value means higher accuracy, so we can say that the model is well trained. In the list ‘value’ we can also see the no. of text entries for each category.
- From that two branches have been made. One is for the negative ones where the team is further more lower (about 0.203) resulting a fine tuned model with a decent accuracy.
- The second branch is for the positive ones where also team index is also lower and predicts better than any other machine learning algorithm.

Potential Improvements:

- **Accuracy improvements:** This is a small dataset. But for large datasets this accuracy should have decreased and further more preprocessing should be done in order to improve accuracy.
- **Hyperparameter Tuning:** Experimenting with different hyperparameter values might lead to better performance.
- **Lack of Deep Learning Model:** Deep learning model is not used in this project. If used then we can say even the percentage of positiveness or negativeness of texts.

```python
from textblob import TextBlob

def get_sentiment(text):
    sentiment_score = TextBlob(text).sentiment.polarity
    if sentiment_score > 0:
        return "Positive"
    elif sentiment_score < 0:
        return "Negative"
    else:
        return "Neutral"

# Example: Apply sentiment analysis to x_test
aspects_sentiment = [get_sentiment(text) for text in x_test]  # Apply function to each text

print("Aspect-Based Sentiment Analysis Results:")
for i, (text, sentiment) in enumerate(zip(x_test, aspects_sentiment)):
    print(f"{i+1}. Aspect: {text} | Sentiment: {sentiment}")
```

```
Aspect-Based Sentiment Analysis Results:
1. Aspect: got girlfriend christma present hasnt abl put | Sentiment: Neutral
2. Aspect: love book reader easi use easi load book easi read even dark easi eye like read real book charg last week | Sentiment: Positive
3. Aspect: great first tablet someon inexpens everyth want | Sentiment: Positive
4. Aspect: bought primarili use eread got cyber monday pretti good deal live googl ecosystem load googl app buy pretti much ignor amazon stuff
5. Aspect: love amazon tv work great need also enjoy play game | Sentiment: Positive
6. Aspect: durabl work great enjoy use book audio book | Sentiment: Positive
7. Aspect: excel product everyth expect perform well voic recognit excel great peopl age 2 year old 90 | Sentiment: Positive
8. Aspect: funsometim difficult phrase question fr alexa love still difficulti find ap enabl differ thing n love music sound good | Sentiment:
9. Aspect: devic good look starter tablet young individu | Sentiment: Positive
10. Aspect: realli like work well easi use | Sentiment: Neutral
11. Aspect: love alexa control thermostat light switch product definit worth money | Sentiment: Positive
12. Aspect: like howev read page number shown also unabl use dictionari still new stage | Sentiment: Positive
13. Aspect: christma gift andmi son pleas | Sentiment: Neutral
14. Aspect: tablet pretti cool reason price easi enough kid use would recommend interest buy one dont wanna spend lot | Sentiment: Positive
```

4. DISCUSSIONS

4.1. STRENGTHS OF THE MODEL:

- **High Accuracy:** The model demonstrated a high level of accuracy in classifying text sentiments as positive, negative or neutral.
- **Robustness:** The model's performance was consistent across different text samples, indicating its robustness.

- **Efficient Training:** The model converged relatively quickly, suggesting efficient learning.

4.2. LIMITATIONS OF THE MODEL:

- **Overfitting Potential:** While the current model shows promising results, there is a risk of overfitting, especially with larger and more complex datasets.
- **Limited Dataset:** The model's performance might be restricted by the size and diversity of the dataset used for training.

4.3. POTENTIAL IMPROVEMENTS:

- **Accuracy improvements:** This is a small dataset. But for large datasets this accuracy should have decreased and further more preprocessing should be done in order to improve accuracy.
- **Hyperparameter Tuning:** Experimenting with different hyperparameter values might lead to better performance.
- **Lack of Deep Learning Model:** Deep learning model is not used in this project. If used then we can say even the percentage of positiveness or negativeness of texts.

4.4. FUTURE WORK

Future research can focus on the following areas:

- **Deep Learning implementation:** Using deep learning model to further improve the accuracy and predicting more precisely.
- **Explainable AI:** Investigating techniques to understand the model's decision-making process.
- **Social Media Validation:** Integrating the model into a social media setting for real-world evaluation and refinement.

5. CONCLUSION

- The developed Sentiment Analysis model has shown promising results in accurately classifying text sentiments.

- While the model exhibits strengths in terms of accuracy and efficiency, addressing potential limitations such as data imbalance and overfitting can further enhance its performance.

- Future research directions include exploring multi-class classification, and explainability to improve the model's social utility.

6. CONCLUSION

- The developed Sentiment Analysis model has shown promising results in accurately classifying text sentiments.

- While the model exhibits strengths in terms of accuracy and efficiency, addressing potential limitations such as data imbalance and overfitting can further enhance its performance.

- Future research directions include exploring multi-class classification, and explainability to improve the model's social utility.