UMT PRESS

Journal of Mathematical Sciences and Informatics
Journal Homepage: https://journal.umt.edu.my/index.php/jmsi
eISSN: 2948-3697
DOI: http://doi.org/10.46754/jmsi.2025.12.005

JMSI Journal of Mathematical Sciences and Informatics

DESIGNING AND IMPLEMENTING AN OPINION MINING ANALYSIS USING ARTIFICIAL INTELLIGENCE

OZOH PATRICK[1]*, IBRAHIM MUSIBAU[1], OYINLOYE OLUFUNKE[1], GBOTOSHO AJIBOLA[1] AND OJO RIDWAN[2]

[1]*Faculty of Computer and Information Technology, Osun State University, Osogbo, Nigeria.* [2]*Faculty of Science, Engineering, and Technology, Osun State University, Osogbo, Nigeria.*

**Corresponding author: patrick.ozoh@uniosun.edu.ng*

ARTICLE INFO

Article History:
Received 11 January 2025
Revised 22 September 2025
Accepted 4 October 2025
Published 15 December 2025

Keywords:
Behavioural analysis;
natural language processing;
online social media;
sentiment analysis;
supervised learning.

ABSTRACT

This study examines the impact of integrating, evaluating, executing, and analysing a model. This involves downloading Twitter information and inserting it into the MongoDB database. The Twitter samples and the extracted features, together with a trained classifier based on supervised learning, their polarity, and emotional words. The insights from the study will help in understanding sentiment analysis using machine learning techniques. The MongoDB database driver, data preprocessing, and sentiment analysis were successfully connected to retrieve text. Visualisations were successful. The application can display graphs and bar charts.

2020 Mathematics Subject Classification:

Introduction

Due to the prominence of social media in contemporary culture, its use has increased the prevalence of illegal activities. This study will help businesses gain insight into their products and services [1]. Several studies on attitudes and feelings related to social networking sites have been conducted, covering the global market, which is shown to be unpredictable. Strunk and White [2] present the impact of various reports on the share price of different companies. Molabahrami [3] explores improvements received in using machine learning techniques. This study presents the results of a combined set of techniques for review in three analysis domains. The domains considered include: (1) Propagators, (2) textual content, and (3) social impact. This system provides insight, explores, and provides perspectives and an overview of future research directions. Liao [4] investigates market returns in the stock market. The study utilises a dataset spanning 10 years. The limitations identified in this study include a major dependence on user metadata, which is insufficient for detection. Furthermore, attributes of profiles alone are inefficient in predicting whether a user is a user of non-credible content.

Zare *et al*. [5] apply machine learning tools for predictions and the influence of predictions on investors' sentiment methods. The drawbacks of this research include the difficulty of exploring investor sentiment due to the diverse platforms available for expression. The obstacles inherent in monitoring offer challenges in interpretation. Sharma *et al*. [6] focus on utilising and improving the techniques. The limitation of this study is that it uses secondary data to analyse the technology industry. This limitation prevents the study from examining trends in real-time applications across other sectors.

This study evaluates online content in terms of numbers and volumes to form an opinion on products. The study methodology is described in Section 3. The enhancements of this research are:

1. Insight into the news or rumours circling on the internet.
2. To combine artificial intelligence and financial modelling in analysing different impacts of sentiments on social media.

Related Works

This section illustrates similar works related to analysing sentiments in opinion mining using machine learning. Opinion mining has aided the status of social media. Jha and Verma [7] apply an analytic model to investigate the effect of organisations' sustainability behaviours using a historical model dependent on information for organisational sustainability performance. The study has a crucial organisational influence on how organisations use social media for sustainability. The study presents achievable insights for the organisation, as well as investors and customers. Albino [8] analysed opinion mining using classification algorithms to evaluate the sentiments of Twitter users. The datasets were collected, and the results were predicted using machine learning techniques. The result indicates an effective structure utilising classification algorithms to aid in evaluating various emotions expressed by social media users.

Archival data on social media dynamics are available online [9]. Social media are collaborative platforms that enable the distribution of information among people. This study utilises image classification to address challenges in information classification. The information available sometimes contains offensive phrases or malicious content that attacks users. These malicious phrases consist of memes, which sometimes integrate an image with a phrase. This study utilises image classification to address challenges in meme classification with current methods and has discovered that selecting troll memes with an image classifier is not feasible. The data gathered from Twitter users is readily used for opinion mining [10]. The study investigates depressive disorder, as a case study, as one of the most trivialised diseases. This behaviour leads to a high proportion of suicide figures worldwide. People with this behaviour can be assessed on their traits on social media platforms to identify individuals who suffer from these issues. This study investigates several methodologies. The objective of this study is to survey methods for investigating profiles that are susceptible to suicide so that appropriate assistance can be given.

The classification of information received from Twitter users is processed for opinion mining [11]. The social media platforms in this study are blogging platforms that provide information for breaking news. As a result, it is crucial to develop effective algorithms to filter through information posted on these platforms and remove unwanted content. This study examined the accuracy of the tweets and their importance.

Rosenthal and Airoldi [12] proposed a technique for automatically processing opinions of Twitter users. The objective of the study [13] is to investigate the occurrence of computer vision syndrome and to evaluate associated factors. The study proposed utilising machine learning to process information on social media platforms. The methods include a descriptive analysis of participants, consisting of students and teachers, with questionnaires administered on social media platforms. The study will provide crucial health information to online users, communicated in a language.

Chhutlani *et al*. [14] applied sentiment analysis to classify information for social media users is based on primary information people look for when making their main decisions. By describing sentiments in association with vital descriptions, the study aims to obtain insights into the concept. By introducing sentiment analysis, valuable insights into specific aspects of operations can be collected, including improvement, which, as a result, allows for a comprehensive experience. The issues of rumour circulation on the Internet and their analysis were presented by [15]. This study proposes a unique standard dataset developed for social media posts. The dataset was created with physical data gathering across different social media platforms, enabling an extensive illustration of user sentiments. This study presents its efficacy in gathering sentiments in social media posts, highlighting a superior dataset standard and adaptability for sentiment analysis in low-quality contexts.

Comparative Analysis of Previous Works

This study compared the techniques reviewed. The study includes comparisons of various tools on various platforms. The reasons for selecting these methods are highlighted, together with the advantages and disadvantages of each technique. These are summarised in the table below to identify the reasons for choosing the methods used in this study.

Table 1: Characteristics of methods

Method	Advantages	Disadvantages
Artificial Neural Network (ANN)	1. The performance evaluation is less complex 2. Reduced overfitting 3. Takes care of missing data 4. Useful for classification and regression functions 5. ANN gives knowledge to features that are important for predictions 6. ANN has performed well using a similar data type	1. Requires much memory and needs many resources 2. Less interpretable than individual decision trees
Naive Bayes	1. Develops training instances for a general model 2. Approximations are made to the target function 3. These algorithms can easily adapt to new data collected over time 4. The Naive Bayes algorithm has been tested on online data	1. A form of lazy learning 2. Relies on storing data 3. The complexity of the hypothesis can grow with the data 4. Each query consists of beginning the new model from scratch, leading to high classification costs
Support Vector Machine (SVM)	1. SVM is adequate for high-dimensional spaces and image classification analysis 2. Handles non-linear relationships 3. Improves accuracy 4. SVM is adequate for text classification	1. Overfitting of models 2. Slow for big datasets, impacting performance 3. Adjusting parameters needs careful tuning 4. SVM is difficult with noisy datasets, reducing its effectiveness

The investigations in this study indicate that the most accurate technique is the ANN, widely applied to social media problems.

Methods

The methods used in this study are described in this section. They include details on data collection and data classification.

Data Gathering

The dataset includes information from various social media platforms processed for information (Figure 1). The dataset collected includes processed tweets made by users. A data classification of the tweets was done to compute the sentiment ratings of the study. The data set also consists of the emotions and opinions of users.

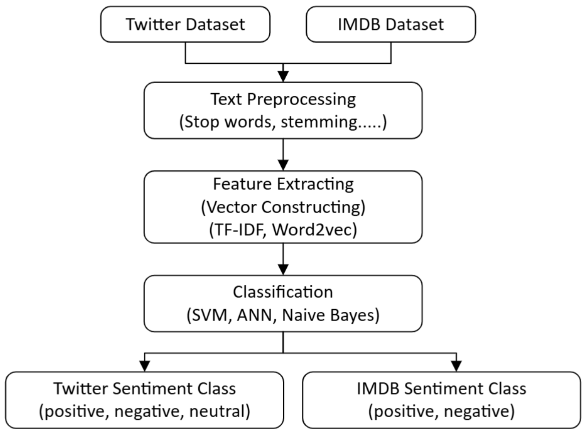

Figure 1: The flowchart

Data Classification

This study classifies data in Twitter posts into various groups and expunges irrelevant information, integrating the approach with natural language processing. This will help in comprehending public views and evaluating user feedback. This is illustrated in Figure 2.

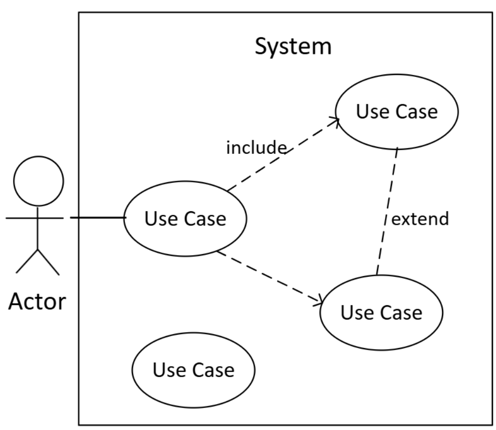

Figure 2: Unified modified language

Description of Algorithms

This entails a description of the models.

Bayes Theorem

This is as given by [16] and presented in Figure 3.

Likelihood

Class Prior Probability

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

Posterior Probability

Predictor Prior Probability

Figure 3: As in Bayes theorem

Maximum Entropy

This technique is described in Equation (1) [17].

$$L\tilde{p}(p) = \log \prod_{x \in X} p(x)^{\tilde{p}(x)} = \sum_{x \in X} \tilde{p}(x) \log p(x) \tag{1}$$

Support Vector Machine (SVM)

The SVM is described in Equation (2) [18].

$$TS_i = \max\left\{\left|\frac{\bar{x}_{ik} - \bar{x}_i}{m_k s_i}\right|, \quad k = 1,2,\ldots,K\right\}$$

$$\bar{x}_{ik} = \sum_{j \in C_k} \bar{x}_{ij} / n_k$$

$$\bar{x}_i = \sum_{j=1}^{n} x_{ij} / n$$

$$s_i^2 = \frac{1}{n-K} \sum_k \sum_{j \in C_k} \left(x_{ij} - \bar{x}_{ik}\right)^2$$

$$m_k = \sqrt{\frac{1}{n_k} + \frac{1}{n}} \tag{2}$$

K-Means Clustering

The algorithm is given as in Equation (3) [19].

$$\mathrm{E(m1,. . ., mM)} = \sum \mathrm{i} = \mathrm{1N} \sum \mathrm{k} = \mathrm{1MI(x} \ldots \tag{3}$$

The Term Frequency Technique

This is as described by [20]. The technique involves allocating weight.

$$W_{i,j} = TF_{t,d}\left(\frac{N}{D_t}\right) \tag{4}$$

Prediction of Sentiments

1. Share positive sentiments.
2. Spread a negative vaccination topic.

The program is listed in Appendix A [21].

Results and Discussions

Implementation

The descriptions for various executions, including the use of the program attached in the appendix [21] are listed in this section. This will help organisations gain insight into the opinions of users, displayed in Figure 4.

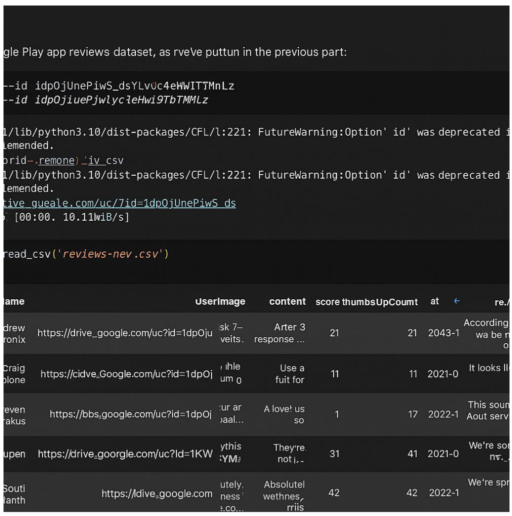

Figure 4: Data exploration

The machine learning model for preprocessing does not work with raw text (Figure 5). The text is converted to numbers. Tokens are added to separate sentences and for classification. A pretrained Bert Tokenizer is then loaded. This text provides an avenue to understand the tokenisation process.

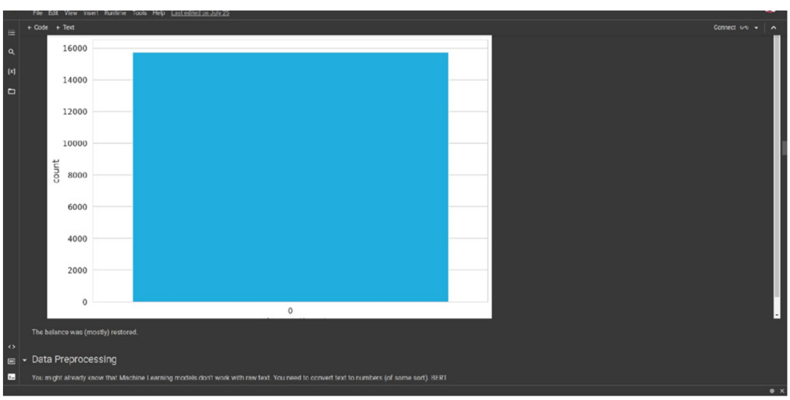

Figure 5: Data preprocessing

The sentiment presents a model for handling noisy data, words, and contextual text (Figure 6). The model extracts contextual semantic features from the embedded data. The figure shows a Jupyter notebook showing the technique. It examines the words in a sentence bidirectionally.

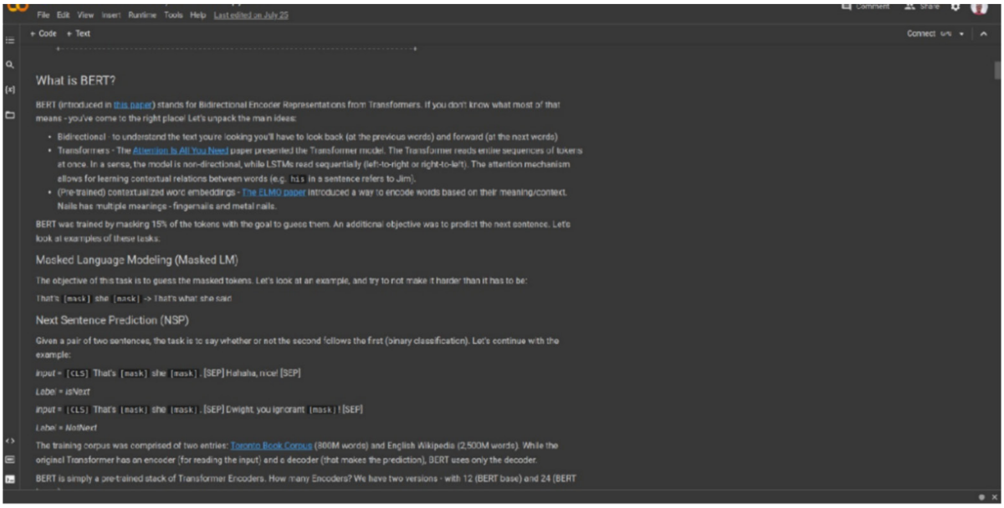

Figure 6: Bidirectional Encoder Representations from Transformers (BERT) representations

The model predicts the text to be encoded, reshaping it into a sequence of tokens and text as numbers, which the model utilises. Each article comprises a list of words (Figure 7).

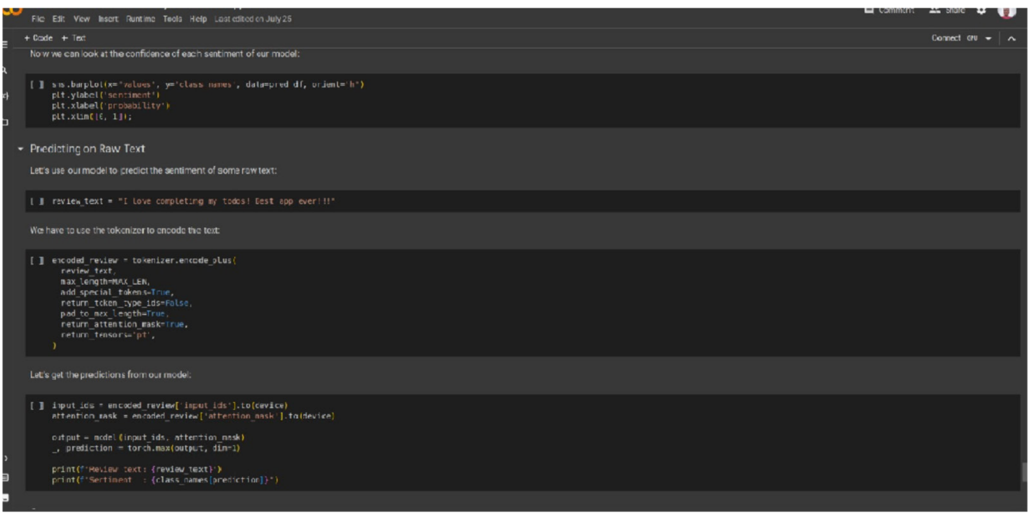

Figure 7: Predicting on raw text

The predicted probabilities are obtained from the trained model by moving the training data to the BERT. In training, a batch is fed to the model to generate the training loop as indicated by the highest value (Figure 8).

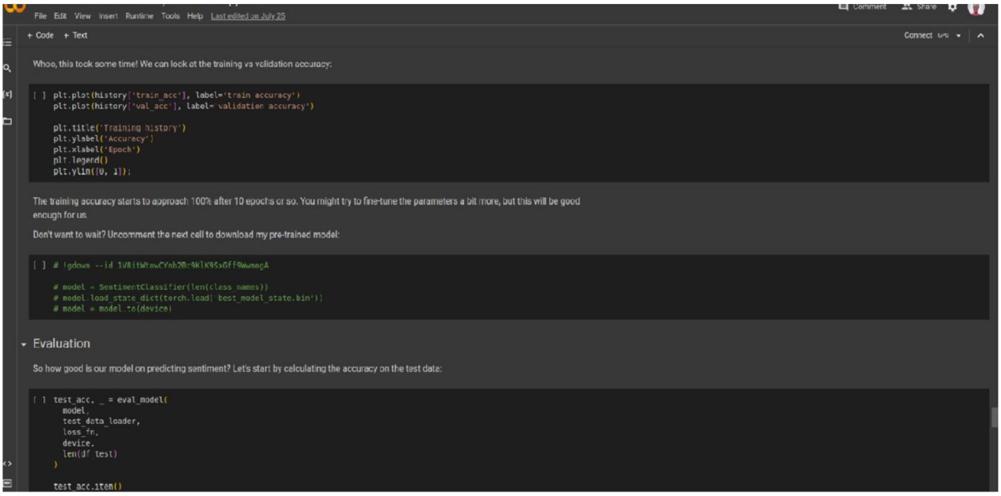

Figure 8: Validation of the study

Figure 9 shows the preprocessing of data, while neutral reviews are in Figure 10.

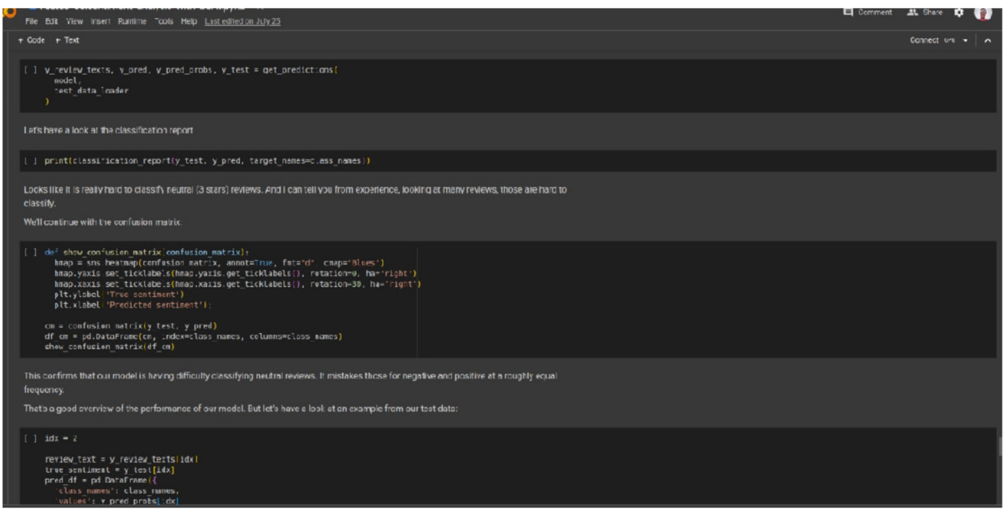

Figure 9: Preprocessing of data

Confirmation of the classification of the reviews is in Figure 10.

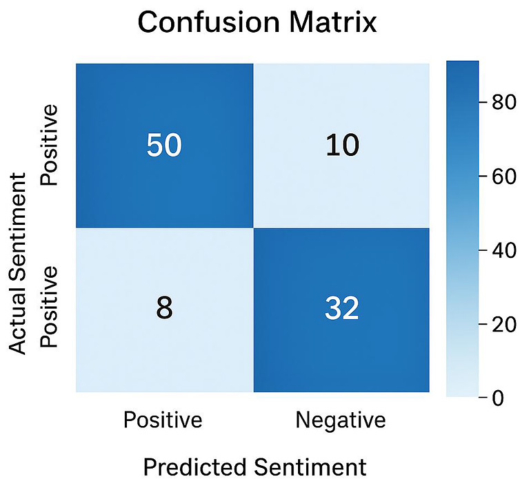

Figure 10: Confusion matrix

To investigate the validity of the model, the accuracy of the data is calculated (Figure 11). The accuracy is 1% lower on test data. This shows the model is accurate.

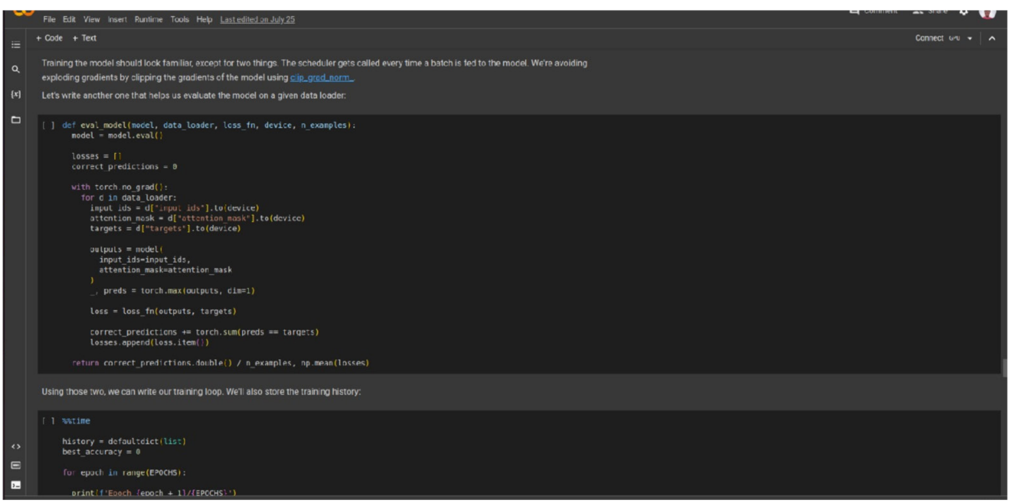

Figure 11: Model evaluation

Interpretations of the Results

The proposed model is tested on the test data (Figure 12). The figure shows a good model. It provides pretrained models and tokenisers for training and evaluation.

```
ns.barplot(x='values', y=class_names', data=pred_df, orient='h')
lt.ylabel('sentiment')
lt.xlabel('probability')
lt.xlim([0, 3]);
```

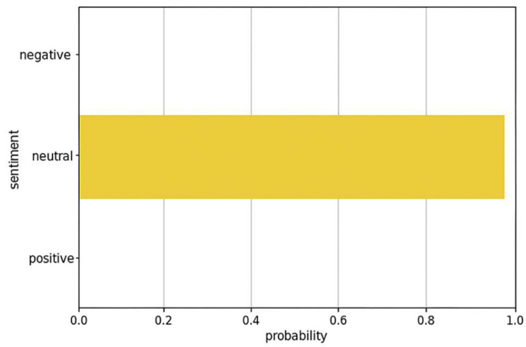

Figure 12: Performance of the proposed model

Conclusions

A sentiment analysis model was developed to retrieve text, including the display of the various visualisations. Furthermore, the proposed model demonstrates good performance and is capable of displaying sentiment analysis results on web applications. Overall, this study has addressed its research objectives.

1. To gain knowledge of tweets
2. To combine artificial intelligence and financial modelling in analysing different impacts of sentiments on social media.

However, this study has certain limitations. They include: (1) Context-dependent errors: The use of sarcasm can be a way of showing negative sentiment, but the words used can be classified as positive. (2) Polarity: Some words cannot be easily classified as positive, negative, or neutral. (3) Negation detection: Some words having negation (e.g., no, not, non-, -less, dis-) do not mean that the sentiment of the statement is negative. A present negation is not enough to classify the sentiment incorrectly. (4) Potential biases in model training: Artificial intelligence algorithms can be prone to human error. There could be a human error in training the dataset. The connection to the database and downloading of tweets were successful, and the results for the study were reliable. The application was executed as specified in the system requirements.

This study proposes the integration of financial modelling with sentiment analysis, leveraging artificial intelligence for the analysis of predictive accuracy in investor sentiment. The use of artificial intelligence improves the model's strength and facilitates a deeper understanding of how investors perceive their actual financial situations. The findings give a practical demonstration of how analysts seek data-driven insights into market decision-making. Future research will help in enhancing opinion mining using sentiment analysis.

Acknowledgement

The authors are grateful to the anonymous reviewers for their helpful comments on improving the article.

Conflict of Interest Statement

The authors declare that they have no conflict of interest.

References

[1] Vyavahare, S., Teraiya, S., & Kumar, S. (2023). FDM manufactured auxetic structures: An investigation of mechanical properties using machine learning techniques. *International Journal of Solids and Structures*, *265*, 112126.

[2] Song, C., Chen, S., Cai, X., & Chen, H. (2024). *Sentiment Analysis of Spanish Political Party Tweets Using Pre-trained Language Models*. ArXiv.org. https://arxiv.org/abs/2411.04862

[3] Plikynas, D., Rizgelienė, I., & Korvel, G. (2025). Systematic review of fake news, propaganda, and disinformation: Examining authors, content, and social impact through machine learning. *IEEE Access, 13, 17583–17629. https://doi.org/10.1109/access.2025.3530688.*

[4] Nguyen, H. H., Ngo, V. M., Pham, L. M., & Van Nguyen, P. (2025). Investor sentiment and market returns: A multi-horizon analysis. *Research in International Business and Finance*, *74*, 102701.

[5] Zare, Z., Sifat, A. I., & Karatas, M. (2025). A review of data analytics and machine learning for personalization in tech sector marketing. *Journal of Soft Computing and Decision Analytics*, *3*(1), 92-111.

[6] Sharma, N. A., Ali, A. S., & Kabir, M. A. (2025). A review of sentiment analysis: Tasks, applications, and deep learning techniques. *International Journal of Data Science and Analytics*, *19*(3), 351-388.

[7] Jha, A. K., & Verma, N. K. (2023). Social media sustainability communication: An analysis of firm behaviour and stakeholder responses. *Information Systems Frontiers*, *25*(2), 723-742.

[8] Albino, M. G. (2024). Opinion-mining technique on Generative Artificial Intelligence topic using data classification algorithms. *Engineering and Technology Journal*, *9*(7).

[9] Suryawanshi, S., Chakravarthi, B. R., Verma, P., Arcan, M., McCrae, J. P., & Buitelaar, P. (2020, May). A dataset for troll classification of TamilMemes. In *Proceedings of the WILDRE5–5th Workshop on Indian Language Data: Resources and Evaluation* (pp. 7-13).

[10] Godara, R. K., Mengi, A., Sharma, A., & Sharma, S. (2023, December). Detecting depressive symptoms on social media: A comprehensive review of methodologies and strategies for suicide prevention. In *International Conference on Cognitive Computing and Cyber Physical Systems* (pp. 87-100). Singapore: Springer Nature Singapore.

[11] Rana, R. S., Jain, K., & Hema, M. (2024, July). Information retrieval on disaster tweets using NLP. In *AIP Conference Proceedings* (Vol. 3075, No. 1). AIP Publishing.

[12] Rosenthal, B., & Airoldi, M. (2024). Consumer morality formation on social media platforms: The case of guns in Brazil. *Journal of Micromarketing*, *44*(1), 178-198.

[13] Chowdhury, S. H., Pathan, M., Arefin, K. E., & Noorul Huq, D. M. (2024). Prevalence and effect of computer vision syndrome during COVID-19: Among Bangladeshi people. *SAS Journal of Medicine*, *4*, 204-211.

[14] Chhutlani, K., Takrani, V., Motwani, A., Harchandani, T., & Sahu, S. (2023, July). Sentiment analysis of OYO Hotel reviews using NLP. In *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)* (pp. 1-6). IEEE.

[15] Mahmud, T., Karim, R., Chakma, R., Chowdhury, T., Hossain, M. S., & Andersson, K. (2024). A benchmark dataset for cricket sentiment analysis in bangla social media text. *Procedia Computer Science*, *238*, 377-384.

[16] Bharadiya, J. P. (2023). A review of Bayesian machine learning principles, methods, and applications. *International Journal of Innovative Science and Research Technology*, *8*(5), 2033-2038.

[17] Patel, D., Saxena, S., Verma, T., & Student, P. G. (2016). Sentiment analysis using maximum entropy algorithm in big data. *International Journal of Innovative Research in Science, Engineering and Technology*, *5*(5).

[18] Wang, L. (Ed.). (2005). *Support vector machines: Theory and applications* (Vol. 177). Springer Science & Business Media.

[19] Likas, A., Vlassis, N., & Verbeek, J. J. (2003). The global k-means clustering algorithm. *Pattern Recognition*, *36*(2), 451-461.

[20] Naeem, M. Z., Rustam, F., Mehmood, A., Ashraf, I., & Choi, G. S. (2022). Classification of movie reviews using term frequency-inverse document frequency and optimized machine learning algorithms. *PeerJ Computer Science*, *8*, e914.

[21] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., & Rush, A. (2020, October). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: System demonstrations* (pp. 38-45).

[22] Acar, Y., Demet, G., & Kaynar, M. (2023). Data and discourse: An assessment of Taksim urban design competition in terms of populism and participation. *Journal of Urban Design*, *28*(6), 682-698.

[23] Nizam, K., & Ahsan, N. (2023). The impact of e-determinants on customer loyalty: A social sustainability paradigm. *Journal of Research in Social Development and Sustainability*, *2*(2), 87-115.

[24] Qian, K., Belyi, A., Wu, F., Khorshidi, S., Nikfarjam, A., Khot, R., Sang, Y., Luna, K., Chu, X., Choi, E., Govind, Y., Seivwright, C., Sun, Y., Fakhry, A., Rekatsinas, T., Ilyas, I., Qi, X., & Li, Y. (2023). Open domain knowledge extraction for knowledge graphs. arXiv preprint arXiv:2312.09424.

Appendix A [21]

Implementation Code

```
!nvidia-smi
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 525.105.17   Driver Version: 525.105.17   CUDA Version: 12.0     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|                               |                      |               MIG M. |
|===============================+======================+======================|
|   0  Tesla T4            Off  | 00000000:00:04.0 Off |                    0 |
| N/A   55C    P8    10W /  70W |      0MiB / 15360MiB |      0%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                                  |
|  GPU   GI   CI        PID   Type   Process name                  GPU Memory |
|        ID   ID                                                   Usage      |
|=============================================================================|
|  No running processes found                                                 |
!pip install -q -U watermark
!pip install -qq transformers
%reload_ext watermark
%watermark -v -p numpy,pandas,torch,transformers
#@title Setup & Config
import transformers
from transformers import BertModel, BertTokenizer, AdamW, get_linear_schedule_with_warmup
import torch
import numpy as np
import pandas as pd
import seaborn as sns
from pylab import rcParams
import matplotlib.pyplot as plt
from matplotlib import rc
from sklearn.model_selection import train_test_split
from sklearn.metrics import confusion_matrix, classification_report
from collections import defaultdict
from textwrap import wrap
from torch import nn, optim
from torch.utils.data import Dataset, DataLoader
import torch.nn.functional as F
%matplotlib inline
```

```
%config InlineBackend.figure_format='retina'
sns.set(style='whitegrid', palette='muted', font_scale=1.2)
HAPPY_COLORS_PALETTE = ["#01BEFE", "#FFDD00", "#FF7D00", "#FF006D", "#ADFF02", "#8F00FF"]
sns.set_palette(sns.color_palette(HAPPY_COLORS_PALETTE))
rcParams['figure.figsize'] = 12, 8
RANDOM_SEED = 42
np.random.seed(RANDOM_SEED)
torch.manual_seed(RANDOM_SEED)
device = torch.device("cuda:0" if torch.cuda.is_available() else "cpu")
device
!gdown --id 1S6qMioqPJjyBLpLVz4gmRTnJHnjitnuV
!gdown --id 1zdmewp7ayS4js4VtrJEHzAheSW-5NBZv
df = pd.read_csv("reviews.csv")
df.head() sns.countplot(df.score)
plt.xlabel('review score');
def to_sentiment(rating):
  rating = int(rating)
  if rating <= 2:
    return 0
  elif rating == 3:
    return 1
  else:
    return 2
df['sentiment'] = df.score.apply(to_sentiment)
```