

American International University-Bangladesh (AIUB)

# Hybrid Model Implementation for Sentiment Analysis on Amazon Data

Rahaman, Hasibur (19-40823-2)

Sithi, Sharannaya Dey (19-40811-2)

Rahman, AKM Shahriyar (19-40787-2)

*A Thesis submitted for the degree of Bachelor of Science (BSc) in Computer Science and Engineering (CSE) at*

*American International University Bangladesh in October,2024*

Faculty of Science and Technology (FST)

## Abstract

*(All candidates must edit this page )*

The abstract should outline the main approach and findings of the thesis / software project and normally must be between **300** and **800** words.

An abstract is a summary of your entire thesis and should provide a complete overview of the thesis, including your key results and findings.

An abstract is different to your introduction, and shouldn’t be used to advertise your thesis — it should provide enough information to allow readers to understand what they’ll learn by reading the thesis.

Your abstract should answer the following questions:

1. What did you do?
2. How did you do it?
3. Why was it worth doing?
4. What were the key results?
5. What are the implications or significance of the results?

As your abstract will have a word limit, you may be unable to answer every question in detail. If you find yourself running out of words, make sure you include your key findings before other information.

## Declaration by author

This thesis is composed of our original work, and contains no material previously published or written by another person except where due reference has been made in the text. We have clearly stated the contribution of others to our thesis as a whole, including statistical assistance, survey design, data analysis, significant technical procedures, professional editorial advice, financial support and any other original research work used or reported in our thesis. The content of our thesis is the result of work we have carried out since the commencement of Thesis.

We acknowledge that copyright of all material contained in my thesis resides with the copyright holder(s) of that material. Where appropriate we have obtained copyright permission from the copyright holder to reproduce material in this thesis and have sought permission from co-authors for any jointly authored works included in the thesis.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Rahaman, Hasibur  (19-40823-2)  Computer Science & Engineering  Faculty of Science & Technology | |  | Sithi, Sharannaya Dey  (19-40811-2)  Computer Science & Engineering  Faculty of Science & Technology | |
|  | | | | |
|  | Rahman, AKM Shahriyar  (19-40787-2)  Computer Science & Engineering  Faculty of Science & Technology | | |  |

## Approval

The thesis titled **“Hybrid Model Implementation for Sentiment Analysis on Amazon Data”** has been submitted to the following respected members of the board of examiners of the department of computer science in partial fulfilment of the requirements for the degree of Bachelors of Science in Computer Science on (**01.10.2024**) and has been accepted as satisfactory.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Tohedul Islam  Assistant Professor, Faculty  Department Of Computer Science  American International University-Bangladesh | |  | S M Abdullah Shafi  Lecturer, Faculty  Department Of Computer Science  American International University-Bangladesh | |
|  | | | | |
|  | Name of the Head (UG/PG)  Rank & Position | | |  |
| Prof. Dr. Dip Nandi  Associate Dean  Faculty of Science and Technology  American International University-Bangladesh | |  | Mashiour Rahman  Sr. Associate Professor & Dean-in-charge  Faculty of Science and Technology  American International University-Bangladesh | |

## Publications included in this thesis

If you choose to include publications as part of your thesis use this section to detail accepted or in press publication(s) using the standard citation format for your discipline.

Papers submitted for publication and awaiting review should appear in the next section, **Submitted manuscripts included in this thesis**.

If you have not included any of your publications in the thesis then state “No publications included”.

#### Example:

1. [[Dum, 2021](#_bookmark17)] **Your Name**, Co-author 1, and Final Author, Title of your paper, *Journal*, Issue, Number, Year
2. [[Dum, 2021](#_bookmark17)] **Your Name**, Co-author 1, and Final Author, Title of your paper, *Journal*, Issue, Number, Year

## Submitted manuscripts included in this thesis

List manuscript/s submitted for publication here. As described above for **Publications included in the thesis**, on the page immediately preceding the chapter that includes the submitted manuscript, in no more than one (1) page, detail your contribution to the authorship if you are not the sole author.

If you have no submitted manuscripts from your candidature then state “No manuscripts submitted for publication”.

#### Example:

1. [[Dum, 2021](#_bookmark17)] **Your Name**, Co-author 1, and Final Author, Title of your paper, submitted to

*Journal* on 4th June 2018.

## Other publications during candidature

List other publications arising during your candidature using the standard citation format for your dis- cipline. Divide your publications into sub-sections as appropriate in your discipline *e.g.* peer-reviewed papers, book chapters, conference abstracts. Papers submitted for publication and awaiting review are not considered publications and cannot be included in this section.

If you have no publications from your candidature then state “No other publications”.

**Example:**

### Conference abstracts

1. [[Dum, 2021](#_bookmark17)] **Your Name**, Co-author 1, and Final Author, Title of your conference paper,

*Proceedings of Conference*, other details.

### Book chapters

1. [[Dum, 2021](#_bookmark17)] **Your Name**, Co-author 1, and Final Author, Title of your chapter, Book, editor,

*etc.*.

## Research involving human or animal subjects

All reserch involving human or animal subjects requires prior ethical review and approval by an independent review committee. At UQ, the relevant committee for research involving human subjects is the [Human Ethics Unit](https://research.cs.aiub.edu/human-ethics) and the relevant committee for research involving animal subjects is the relevant [Animal Ethics Committee](https://research.cs.aiub.edu/animal-welfare). Please provide details of any ethics approvals obtained including the ethics approval number and name of approving committees. A copy of the ethics approval letter must be included in the thesis appendix.

If no human or animal subjects were involved in this research please state: “No animal or human subjects were involved in this research”.

## Contributions by authors to the thesis

List the significant and substantial inputs made by different authors to this research, work and writing represented and/or reported in the thesis. These could include significant contributions to: the conception and design of the project; non-routine technical work; analysis and interpretation of research data; drafting significant parts of the work or critically revising it so as to contribute to the interpretation.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Rahaman, Hasibur** | **Sithi, Sharannaya Dey** | **Rahman, AKM Shahriyar** | **Contribution (%)** |
|  | *(19-40823-2)* | *(19-40811-2)* | *(19-40787-2)* |
| Conceptualization |  |  |  | 100 % |
| Data curation |  |  |  | 100 % |
| Formal analysis |  |  |  | 100 % |
| Investigation |  |  |  | 100 % |
| Methodology |  |  |  | 100 % |
| Implementation |  |  |  | 100 % |
| Validation |  |  |  | 100 % |
| Theoretical derivations |  |  |  | 100 % |
| Preparation of figures |  |  |  | 100 % |
| Writing – original draft |  |  |  | 100 % |
| Writing – review & editing |  |  |  | 100 % |

## Acknowledgments

We would like to express our gratitude to Tohedul Islam, Assistant Professor, Faculty of the Department of Computer Science, and our thesis advisor, for his unwavering support and guidance. We sincerely appreciate all his advice and encouragement throughout our research.

We are very thankful for all his help, guidance, and motivation during our study. Sir has been an exceptional mentor, and his expertise, dedication, and constructive feedback have been invaluable. Our academic goals and ideas have been greatly aided by his support. We have learned a lot, and used our knowledge and insights to challenge ourselves and strive for greatness. We are extremely fortunate to have had the opportunity to work with someone so extraordinary, and we will be eternally grateful to him for his guidance, kindness, and unwavering dedication to our success.

## Keywords

Hybrid Model, Sentiment Analysis, Amazon Data, Vader, RoBERTa, BERT, CNN, LSTM, SVM, TextBlob

Table of Contents

[Hybrid Model Implementation for Sentiment Analysis on Amazon Data i](#_Toc178382381)

[Abstract ii](#_Toc178382382)

[Declaration by author iii](#_Toc178382383)

[Approval 4](#_Toc178382384)

[Publications included in this thesis v](#_Toc178382385)

[Submitted manuscripts included in this thesis v](#_Toc178382386)

[Other publications during candidature vi](#_Toc178382387)

[Conference abstracts vi](#_Toc178382388)

[Book chapters vi](#_Toc178382389)

[Research involving human or animal subjects vi](#_Toc178382390)

[Contributions by authors to the thesis vii](#_Toc178382391)

[Acknowledgments viii](#_Toc178382392)

[Keywords ix](#_Toc178382393)

[List of Figures xi](#_Toc178382394)

[List of Tables xii](#_Toc178382395)

[List of Abbreviations and Symbols xiii](#_Toc178382396)

[Introduction 14](#_Toc178382397)

[1.1 Motivation 14](#_Toc178382398)

[1.2 Objective 14](#_Toc178382399)

[1.3 Research Questions 14](#_Toc178382400)

[1.4 Document Outline 14](#_Toc178382401)

[Literature review 16](#_Toc178382402)

[2.1 Introduction 16](#_Toc178382403)

[Methods 18](#_Toc178382404)

[3.1 Programming Environment and Technologies 18](#_Toc178382405)

[3.2 Data Collection 19](#_Toc178382406)

[3.3 Data Cleaning and Preprocessing 20](#_Toc178382407)

[3.4 Sentiment Analysis Models 21](#_Toc178382408)

[3.5 Methodology 22](#_Toc178382409)

[Results or findings 30](#_Toc178382410)

[9.1 Introduction 30](#_Toc178382411)

[Discussion 32](#_Toc178382412)

[5.1 Introduction 32](#_Toc178382413)

[Conclusion 34](#_Toc178382414)

[Bibliography 36](#_Toc178382415)

[Appendix 38](#_Toc178382416)

[A.1 Name of Appendix-1 38](#_Toc178382417)

[Example of Citations 41](#_Toc178382418)

[Example of Equations 44](#_Toc178382419)

[Example of Figures 46](#_Toc178382420)

[Example of Tables 48](#_Toc178382421)

[Example of algorithm procedure 51](#_Toc178382422)

[Example of Code 54](#_Toc178382423)

[G.1 Find the greatest number from a list of numbers in *Python* 54](#_Toc178382424)

# List of Figures

[Figure D. 1 : American International University-Bangladesh (AIUB) 36](#_Toc76225927)

[Figure D. 2: Example of placing images side by side 38](#_Toc76225928)

# List of Tables

[Table E. 1: AIUB currently operates under four distinct Faculties x 41](#_Toc76193928)

[Table E. 2: Row span example 41](#_Toc76193929)

[Table E. 3 : Test Table 41](#_Toc76193930)

# List of Abbreviations and Symbols

|  |  |
| --- | --- |
| Abbreviations | |
| VADER | Valence Aware Dictionary and sEntiment Reasoner |
| RoBERTa | Robustly Optimized BERT Pretraining Approach |
| BERT | Bidirectional Encoder Representations from Transformers |
| LSTM | Long Short-Term Memory |
| SVM | Support Vector Machine |
| CNN | Convolutional Neural Network |
| POS | Part of Speech (tagging) |
| TF-IDF | Term Frequency-Inverse Document Frequency |

**Chapter 1**

# Introduction

It's now common and popular practice to read online customer reviews and ratings before making a purchase. Based on feedback, consumers are more likely to purchase a product. Because online marketplaces have grown in popularity over the past few decades, many online retailers and sellers ask their customers to leave reviews of the goods they have purchased. Millions of reviews are generated every day throughout the Internet regarding various goods, services, and locations. Due to this, the internet is the most reliable resource for ideas and reviews regarding a good or service. Product reviews offer valuable insights into a product's features, quality, and recommendations, helping potential buyers gain a comprehensive understanding of it. These reviews benefit not only consumers but also sellers, particularly those who manufacture their own products, by providing a better understanding of consumer preferences and needs. However, it is getting harder for a customer to decide whether or not to purchase a product as more reviews become available for it. Customers are more confused when trying to make the right decision due to conflicting opinions about the same product and unclear reviews. In this case, it appears that all e-commerce enterprises must analyze this content.

## 1.1 Motivation

The motivation for developing a hybrid sentiment analysis model stems from the need to enhance the accuracy, robustness, and scalability of existing sentiment analysis approaches. Individual models like VADER, RoBERTa, or BERT perform well in specific scenarios but struggle with diverse text formats, such as short tweets versus lengthy product reviews. The hybrid model leverages multiple models to provide a more comprehensive sentiment analysis, resulting in deeper insights, especially for business decision-making. By combining rule-based, machine learning, and deep learning methods, this model aims to overcome the limitations of standalone approaches and improve sentiment prediction in domains like e-commerce and product reviews.

## 1.2 Objective

The primary objective of this research is to develop a hybrid sentiment analysis model that integrates various sentiment analysis techniques (VADER, RoBERTa, BERT, SVM, CNN, LSTM, TextBlob) to:

* Increase sentiment prediction accuracy across diverse datasets.
* Provide actionable insights for businesses, particularly in the context of customer feedback and product improvement.
* Enhance decision-making processes by producing a robust system that can generalize well across various text types and domains (e.g., Amazon product reviews, social media).

## 1.3 Research Questions

How can a hybrid sentiment analysis model improve the overall sentiment prediction accuracy compared to individual models?

Which model combinations are the most effective in capturing sentiment nuances in diverse data types (short vs. long text, formal vs. informal language)?

How can sentiment insights from this hybrid model contribute to better business decision-making, particularly in product and service improvement?

Can a hybrid model be generalized to other domains outside e-commerce for sentiment analysis, such as social media monitoring or market research?

## 1.4 Document Outline

The rest of the document is as follows. Chapter 2 provides a literature analysis of the sentiment analysis as well as relevant research, journals, and articles reviewed for the idea. Chapter 3 explains justification and explanation of the methodological approach. The results from the experiments are gathered in Chapter 4 and discussed in Chapter 5. Finally, Chapter 6 concludes the study.

**Chapter 2**

# Literature review

## Introduction

The Sentiment analysis, also known as opinion mining, is a subfield of natural language processing (NLP) that aims to identify and extract subjective information from text, such as opinions, emotions, and attitudes. Hybrid models, which combine multiple techniques or models, have emerged as a promising approach to improve sentiment analysis performance. This literature review explores recent research on hybrid models for sentiment analysis, focusing on the integration of semantic knowledgebases, machine learning algorithms, and deep learning techniques. The enumerated summaries of various papers within their field of study are provided below.

## VADER

VADER is a lexicon-based sentiment analysis tool specifically designed for social media text. It uses a predefined list of words and their associated sentiment scores, providing a simple yet effective way to gauge sentiment polarity (positive, negative, neutral). VADER's ability to capture sentiment in short, informal texts makes it particularly useful in real-time applications like social media monitoring and customer feedback analysis. Its effectiveness has been demonstrated in various studies, showcasing its role in hybrid models that enhance sentiment classification by providing initial sentiment scores before further analysis (Alfrjani et al., 2019) [1].

## TextBlob

TextBlob is another lexicon-based model that extends the capabilities of VADER by offering a broader range of text processing tools. It simplifies the task of sentiment analysis by providing an API that enables users to extract sentiment scores easily. TextBlob is particularly advantageous in applications requiring rapid sentiment analysis, as it efficiently handles basic sentiment classification tasks. In hybrid models, TextBlob can serve as a preprocessing step, providing foundational sentiment scores to enhance more sophisticated algorithms (Horvat et al., 2024) [8]

## BERT

BERT is a state-of-the-art transformer-based language model that captures the context of words by considering both left and right surrounding words in a sentence. This bidirectional approach allows BERT to understand the nuanced meaning of language, making it highly effective for tasks like sentiment analysis. In hybrid models, BERT is often combined with traditional machine learning classifiers or other deep learning architectures to improve sentiment classification accuracy, particularly in complex, domain-specific texts (Kaur et al., 2023) [12].

## RoBERTa

RoBERTa is an optimized variant of BERT that focuses solely on the pre-training phase, removing certain tasks like next sentence prediction and enhancing the model's performance on text classification tasks. It leverages a larger training dataset and longer training time, resulting in improved sentiment classification capabilities. In hybrid architectures, RoBERTa's robust language modeling is utilized alongside other models to achieve high accuracy in tasks like aspect-based sentiment analysis (Gupta, 2023) [7].

## CNN

CNNs are primarily known for their applications in image processing but have also proven effective in text analysis. In sentiment analysis, CNNs can capture local patterns in the text, such as phrases or n-grams, by applying convolutional filters. This capability allows CNNs to extract significant features from textual data, which can then be classified using various algorithms. Hybrid models combining CNNs with other techniques, such as LSTMs or SVMs, enhance sentiment classification by leveraging both feature extraction and classification strengths (Dang et al., 2021) [6]

## LSTM Networks

LSTMs are a type of recurrent neural network (RNN) designed to model sequential dependencies in data. They are particularly effective in capturing long-range dependencies in text, making them well-suited for sentiment analysis tasks involving longer documents or conversations. When integrated into hybrid models, LSTMs can work alongside CNNs or transformers to effectively analyze sentiment by considering both local and global context in the text (Rehman et al., 2019) [4]

## SVM

SVMs are traditional machine learning classifiers that excel in handling high-dimensional data and are particularly effective in binary classification tasks. In the context of sentiment analysis, SVMs can be used to classify extracted features from textual data. Hybrid models that combine SVMs with deep learning architectures, such as CNNs or transformers, leverage the feature extraction capabilities of the latter while maintaining the robustness of SVMs for classification, leading to improved sentiment classification outcomes (Elleuch et al., 2016) [3].

## TF-IDF

TF-IDF is a feature extraction method that transforms text into a numerical format, emphasizing important terms based on their frequency within a document relative to their occurrence across a corpus. This technique is often integrated into hybrid models to create feature vectors that can be analyzed by machine learning classifiers. By combining TF-IDF with advanced models like BERT or CNNs, hybrid sentiment analysis models can benefit from both shallow and deep textual features, significantly enhancing performance in applications like product reviews (GR Narayanaswamy et al., 2021) [9].

## Hybrid Models

Hybrid NLP models combine various machine learning and deep learning techniques to create more effective sentiment analysis systems. These models often integrate lexicon-based approaches with advanced algorithms, allowing them to handle complex, nuanced texts more effectively. For instance, a hybrid model might combine BERT with domain-specific knowledge to analyze public sentiment during crises, showcasing the flexibility and adaptability of hybrid frameworks in tackling various sentiment analysis challenges (Gupta, 2023) [7].

## Related Works

Several studies have explored hybrid models that integrate semantic knowledge bases with machine learning techniques. Alfrjani et al. (2019) [1] introduced a *Hybrid Semantic Knowledgebase-Machine Learning Approach*, which integrates structured semantic knowledge with machine learning algorithms. This combination improves opinion mining by leveraging external knowledge bases to handle ambiguous or domain-specific sentiments more effectively. By incorporating semantic understanding into machine learning, this model mitigates issues related to ambiguous language and improves classification accuracy.

Elshakankery (2019) [2] developed *HILATSA*, a hybrid incremental learning approach designed for analyzing Arabic tweets. This model adapts dynamically as it processes new data, making it ideal for real-time, large-scale sentiment analysis tasks. The hybrid structure combines incremental learning techniques with traditional sentiment analysis algorithms to enhance adaptability and accuracy.

Elleuch et al. (2016) [3] introduced a *CNN-SVM Hybrid Model* for offline Arabic handwritten recognition, which uses CNN for feature extraction and SVM for classification. While this model was primarily designed for text recognition, its architecture can be adapted to sentiment analysis by utilizing CNN for extracting features from text and SVM for sentiment classification.

Rehman (2019) [4] proposed a *Hybrid CNN-LSTM Model* for movie review sentiment analysis. In this model, CNN is used for spatial feature extraction, while LSTM captures sequential dependencies in the text. This approach effectively combines the strengths of both models, leading to improved accuracy in sentiment classification by leveraging both spatial and temporal features.

Shobayo et al. (2024) [5] conducted a comparative study focusing on customer sentiment analysis using Google PaLM, demonstrating how hybrid architectures that combine pretrained models like Google PaLM with traditional machine learning techniques result in superior performance for product review analysis.

Dang et al. (2021) [6] and Gupta et al. (2023) [7] explored hybrid models that combine CNN, LSTM, and transformers to handle sentiment analysis more effectively. These models leverage CNN’s feature extraction capabilities, LSTM’s strength in sequential data, and transformers’ contextual understanding to improve the classification of complex text datasets.

Horvat (2024) [8] introduced a *Hybrid Natural Language Processing (NLP) Model* for analyzing public sentiments during natural crises. This model integrates BERT with domain-specific knowledge bases to assess public sentiment more accurately. The hybrid model captures the complexities of crisis-related texts, showing significant improvements in context-specific sentiment analysis.

Narayanaswamy et al. (2021) [9] explored the combination of BERT and RoBERTa for aspect-based sentiment analysis (ABSA). The hybrid model leverages BERT’s contextual language understanding and RoBERTa’s robust fine-tuning capabilities to achieve high accuracy in identifying sentiments associated with specific aspects of products or services.

Semary et al. (2023) [10] emphasized the importance of feature extraction in hybrid models for sentiment analysis. By combining traditional feature extraction techniques such as TF-IDF with deep learning architectures like CNN and transformers, they demonstrated how hybrid models can achieve better performance on sentiment classification tasks.

Semary et al. (2024) [11] and Kaur et al. (2023) [12] applied hybrid deep learning models for consumer sentiment analysis, combining CNN, LSTM, and transformer models. These models efficiently capture sentiment variations in consumer product reviews by integrating advanced feature extraction techniques, improving accuracy and performance.

Kaur et al. (2023) [13] proposed a *RoBERTa-based hybrid model* for improving sentiment classification. This hybrid model combines RoBERTa’s deep contextual embeddings with traditional machine learning classifiers, leading to improved performance in sentiment analysis tasks, particularly when handling complex textual data.

Yürütücü et al. (2021) [14] introduced a hybrid model that integrates sentiment scores with readability metrics for filtering spam reviews. Their research highlighted how hybrid models combining sentiment analysis with linguistic features can outperform traditional spam filtering techniques, especially in the context of product reviews.

Kanmani (2023) [15] proposed a multi-channel *LSTM-CNN Hybrid Model* for Vietnamese sentiment analysis. The model uses CNN for feature extraction and LSTM for processing sequential data, resulting in more accurate sentiment classifications. The combination of spatial and sequential processing proved beneficial in handling complex Vietnamese texts.

Aggarwal, et al. (2023) [16] and QH Vo et al. (2027) [17] explored the use of pretrained language models like BERT, GPT, and RoBERTa for sentiment analysis. Their studies demonstrated that combining pretrained models with traditional feature extraction techniques leads to significant improvements in classification accuracy, particularly when handling sentiment nuances in large-scale datasets.

Keith et al. (2017) [18] presented a comprehensive analysis of mobile product reviews using hybrid sentiment analysis models. By integrating deep learning models with aspect-based analysis techniques, their approach yielded superior performance in sentiment classification and aspect identification.

**Chapter 3**

# Methods

This chapter gives the reader an overview of the different tools and procedures used to accomplish the Hybrid Model Implementation for Sentiment Analysis on Amazon Data research. The reader will also become familiar with each technology's basic overview and the design choices made to achieve it. In the first section, the programming environments will be discussed. The second section will cover about the data collection procedure. And then in the third section Data Cleaning Procedure will be explain. The fourth section will explain data splitting procedure. After that on fifth section model building and training procedure will be explained. The entire methodology used for this research will be explained in the final section.

## Programming Environment and Technologies

One of the most popular programming languages for data science and machine learning is Python. Numerous machine learning algorithms can be solved using the extensive library of Python. That’s why we have chosen python as the programming language for this research purpose. Python, libraries that we have used in this project are,

* + 1. Pandas,
    2. re (regular expression)
    3. NLTK (Natural Language Toolkit),
    4. BeautifulSoup
    5. Matplotlib,
    6. Seaborn,
    7. TextBlob
    8. TfidfVectorizer
    9. SVC
    10. numpy
    11. Transformers,
    12. Torch,
    13. TensorFlow,

## Data Collection

Data was collected from amazon website by scrapping using Python code as Amazon does not have an API like Twitter to download reviews with. Later, the dataset was saved in the Comma Separated Values (CSV) format because Python can handle these kinds of files more easily. Datasets that we had collected is consist of Amazon Smartphone review Data. The fields listed below are included in the dataset:

* + 1. Id: A unique identifier
    2. **product\_title:** Name of the product is stored here
    3. **user\_name:** Name of the reviewer is stored in this field
    4. **rating:** rating of the product
    5. **review:** text of the review
    6. **review\_date:** date of the review

The Dataset contains a total of 4,574 reviews.

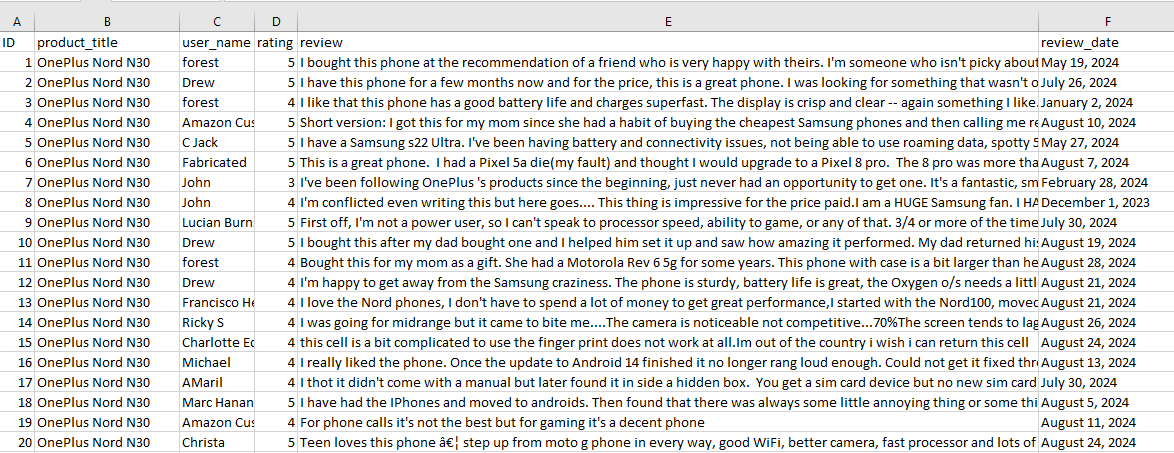


Fig-1: Sample Datasets

## Data Preprocessing

The data preprocessing steps in this model are integral to ensure the text data is in a suitable format for training a neural network model effectively. We have pre process the data in 4 different ways for different models. First, we have created a common process function where we have converted the text into lower case, remove the numbers and extra spaces from the text. Then for VADER we will apply this normal preprocessing directly. For transformers model we have remove the punctuation except the meaningful ones (, ! ?). After that for the Deep Learning models we have removed all the punctuations. And for machine learning model SVM we have removed the punctuations and stopwords. At last, we have assigned the value in different variables and store them into another CSV file.

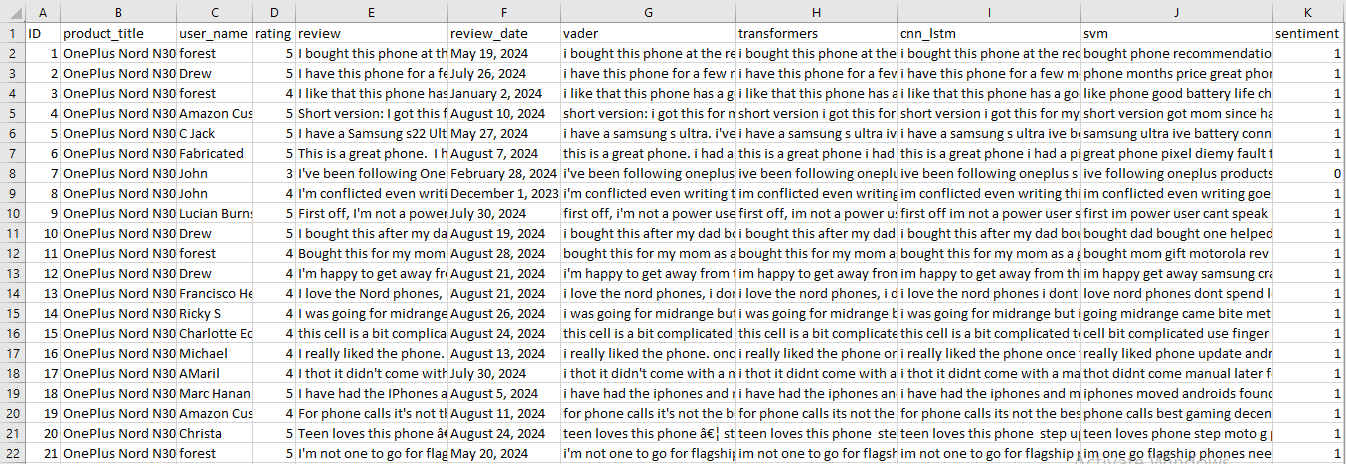


Fig-2: Pre-Processed Data

## Splitting Data

Splitting data is a crucial preprocessing step in machine learning projects, aimed at dividing the dataset into distinct subsets for training and testing. This process ensures that the model is trained on one portion of the data and tested on a separate subset to evaluate its performance on unseen data. Typically, the dataset is divided into two subsets: training and testing, often in ratios such as 70-30 or 80-20. The training set, which contains most of the data, is used to train the model. The test set is executed and utilized to assess the final performance of the trained model.

Splitting the data ensures that the model's performance is evaluated on unseen data, providing an unbiased estimate of its generalization ability. This step helps in detecting overfitting, where the model performs well on the training data but poorly on unseen data and allows for fine-tuning model parameters 29 to optimize performance. Data separation is particularly important for NLP tasks, as it enables the model to generalize patterns in text data and accurately predict new, unknown text. As we have preprocess data in different way so we have split the data into multiple train test portion.

## Model Building

In this section we will describe the process of all the model we have built.

**Chapter 4**

# Results or findings

Use the results section to:

* + 1. specify the data you collected and how it was were prepared for analysis
    2. describe the data analysis (e.g. define the type of statistical test that was applied to the data)
    3. describe the outcome of the analysis
    4. present a summary and descriptive statistics in a table or graph.

#### Use tables and figures effectively

Reports usually include tables, graphs and other graphics to present data and supplement the text. To learn how to design and use these elements effectively, see examples provided in Appendix B ([D](#_bookmark27), [E](#_bookmark30), [C](#_bookmark26), [F](#_bookmark34)).

Introduce the broad layout of the chapter.

## Introduction

Add your text here.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | Accuracy | Precision (Macro) | Precision (Micro) | Precision (Weighted) | Recall | F1-Measure |
| **VADER** | 0.6361 | 0.3216 | 0.6361 | 0.5739 | 0.4272 | 0.3540 |
| **RoBERTa** | 0.6852 | 0.3606 | 0.6852 | 0.6289 | 0.4986 | 0.3979 |
| **BERT** | 0.5607 | 0.3021 | 0.5607 | 0.5673 | 0.3718 | 0.3130 |
| **TextBlob** | 0.6350 | 0.3045 | 0.6350 | 0.5549 | 0.3917 | 0.3360 |
| **SVM** | 0.6885 | 0.3289 | 0.6885 | 0.5909 | 0.4271 | 0.3658 |
| **CNN** | 0.5202 | 0.3545 | 0.5202 | 0.6513 | 0.4635 | 0.3328 |
| **LSTM** | 0.6240 | 0.2949 | 0.6240 | 0.5217 | 0.3833 | 0.3303 |
| **Hybrid** | 0.6765 | 0.3638 | 0.6765 | 0.6359 | 0.5071 | 0.3980 |

**Chapter 5**

# Discussion

Use the discussion section to:

* + - * 1. comment on your results and explain what they mean
        2. compare, contrast and relate your results back to theory or the findings of other studies
        3. identify and explain any unexpected results
        4. identify any limitations to your research and any questions that your research was unable to answer
        5. discuss the significance or implications of your results.
        6. If you find that your research ends up in a different direction to what you intended, it can help to explicitly acknowledge this and explain why in this section.

Introduce the broad layout of the chapter.

## Introduction

Add your text here.

**Chapter 6**

# Conclusion

Use the conclusion section to:

* + 1. summarise the main findings of your research
    2. emphasise that you’ve met your research aims. A good strategy is to repeat your research questions and demonstrate how your findings answer them.
    3. restate the limitations of your research and make suggestions for further research.

In some cases, the discussion and conclusion sections can be combined. Check with your supervisor if you want to combine these sections. your conclusion chapter should not exceed two pages.

Conclude your thesis.

# Bibliography

1. R Alfrjani, T Osman, G Cosma, “A Hybrid Semantic Knowledgebase-Machine Learning Approach for Opinion Mining” - Data & Knowledge Engineering, Elsevier, Volume 121, May 2019, Pages 88-108, doi: <https://doi.org/10.1016/j.datak.2019.05.002>
2. K Elshakankery, MF Ahmed, “A hybrid Incremental learning approach for Arabic tweets sentiment analysis” - Egyptian Informatics Journal, Elsevier, Volume 20, Issue 3, November 2019, Pages 163-171, doi: <https://doi.org/10.1016/j.eij.2019.03.002>
3. M Elleuch, R Maalej, M Kherallah, “A New Design Based-SVM of the CNN Classifier Architecture with Dropout for Offline Arabic Handwritten Recognition”- Procedia Computer Science, Elsevier, Volume 80, 2016, Pages 1712-1723, doi: [https://doi.org/10.1016/j.procs.2016.05.512](https://doi.org/10.1016/j.procs.2016.05.512%20)
4. AU Rehman, AK Malik, B Raza, W Ali, “A Hybrid CNN-LSTM Model for Improving Accuracy of Movie Reviews Sentiment Analysis” - Multimedia Tools and Applications, pringer, Volume 78, page 26597-26613, (2019), doi: <https://doi.org/10.1007/s11042-019-07788-7>
5. O Shobayo, S Sasikumar, S Makkar, O Okoyeigbo, “Customer Sentiments in Product Reviews: A Comparative Study with GooglePaLM” - Analytics, dpi.com,Volume 2024,3(2),241-254, doi: <https://doi.org/10.3390/analytics3020014>
6. CN Dang, MN Moreno-García, F De la Prieta, “Hybrid Deep Learning Models for Sentiment Analysis” - Complexity, Wiley Online Library, Volume 2021, Issue 1/9986920, doi: <https://doi.org/10.1155/2021/9986920>
7. A Gupta, I Arhatia, D Gautam, R Verma, “Mobile Review Sentiments: An In-Depth Analysis” - researchgate.net
8. M Horvat, G Gledec, F Leontić, “Hybrid Natural Language Processing Model for Sentiment Analysis during Natural Crisis” - Electronics, mdpi.com, Volume Electronics 2024, 13(10), 1991 , doi: <https://doi.org/10.3390/electronics13101991>
9. GR Narayanaswamy, “Exploiting BERT and RoBERTa to Improve Performance for Aspect Based Sentiment Analysis” - arrow.tudublin.ie,Volume Narayanaswamy, G.R. (2021), doi: <https://doi.org/10.21427/3w9n-we77>
10. NA Semary, W Ahmed, K Amin, P Pławiak, “ Improving sentiment classification using a RoBERTa-based hybrid model” - Frontiers in Human,frontiersin.org, Volume 17 - 2023, doi: <https://doi.org/10.3389/fnhum.2023.1292010>
11. N A. Semary, W Ahmed, K Amin, P Pławiak, Mohamed Hammad “Enhancing machine learning-based sentiment analysis through feature extraction techniques” - Plos one, journals.plos.org, Volume February 14, 2024 doi: <https://doi.org/10.1371/journal.pone.0294968>
12. G Kaur, A Sharma “HAS: Hybrid Analysis of Sentiments for the perspective of customer review summarization”-Journal of Ambient Intelligence and Humanized, Springer, Volume 14, pages 11971–11984, (2023), doi: <https://doi.org/10.1007/s12652-022-03748-6>
13. G Kaur, A Sharma “A deep learning‑based model using hybrid feature extraction approach for consumer sentiment analysis”-Journal of big data, Springe,Volume 10, article number 5, (2023), doi: <https://doi.org/10.1186/s40537-022-00680-6>
14. ÖY Yürütücü, “THE USE OF PRETRAINED LANGUAGE MODELS IN SENTIMENT ANALYSIS” - search.proquest.com, APRIL 2022,69 Pages, doi:
15. S Kanmani, S Balasubramanian, “Leveraging Readability and Sentiment in Spam Review Filtering Using Transformer Models”- Computer Systems Science & Engineering,cdn.techscience.cn, Vol.45, No.2, 2023 doi: <https://doi.org/10.32604/csse.2023.029953>
16. R Aggarwal, L Gupta, “A Hybrid Approach for Sentiment Analysis using Classification Algorithm”- International Journal of Computer Science and Mobile
17. QH Vo, HT Nguyen, B Le “Multi-channel LSTM-CNN model for Vietnamese sentiment analysis”- 9th international, ieeexplore.ieee.org, Volume 19-21 October 2017, doi: <https://doi.org/10.1109/KSE.2017.8119429>
18. B Keith, E Fuentes, C Meneses, “A Hybrid Approach for Sentiment Analysis using Classification Algorithm” - Proceedings of ACM SIGKDD, blog.sentic.net, Volume August 2017 (KDD’17), 10 pages,

**Appendix A**

# Appendix

Write your appendix here. Following two are examples.

## Name of Appendix-1

* 1. **Name of Appendix-2**

**Appendix B**

# Example of Citations

This text is only for Bibliography testing purposes.

Dr. Dip Nandi currently works as an Associate Professor and the Director of Faculty of Science and Technology in American International University-Bangladesh (AIUB). His research area includes: Software Engineering, E-Learning Technologies, Data Mining and Information systems and has produced several publications in these domains [[Nandi et al., 2012](#_bookmark20), [Nandi et al., 2011](#_bookmark21)].

Dr. Tabin Hasan primarily focuses in the research Domain of Human Computer Interaction. He is been a active researcher for more than a decade and produced many high quality journals [[Hasan et al.,](#_bookmark18) [2013](#_bookmark18)], conferences [[Hasan et al., 2021](#_bookmark19)] and book chapters.

**Appendix C**

# Example of Equations

The well known Pythagorean theorem *x*2 + *y*2 = *z*2 was proved to be invalid for other exponents. Meaning the next equation has no integer solutions:

*xn* + *yn* = *zn*

The ampersand character & determines where the equations align. Let’s check a more complex example:

*x* = *y w* = *z a* = *b* + *c*

2*x* = *y* 3*w* = 1 *z a* = *b*

*−*

2

*−*4 + 5*x* = 2 + *y w* + 2 = *−*1 + *w ab* = *cb*

The mass-energy equivalence is described by the famous equation

*E* = *mc*2

discovered in 1905 by Albert Einstein. In natural units (*c* = 1), the formula expresses the identity

Some random examples ...

*E* = *m* (C.1)

∞ 1 1

∑ *ns* = ∏ 1 *− p−s* (C.2)

*i*=1 *p*

∞

∑ 2*−n* = 1 (C.3)

*n*=1

*V µ*(*t, u, v, w*) *dt dudvdw* (C.4)

**Appendix D**

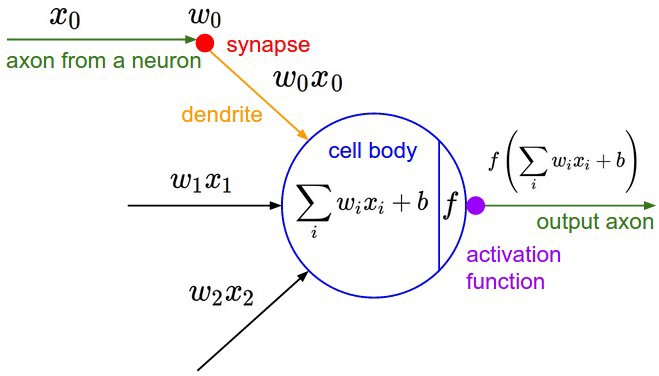
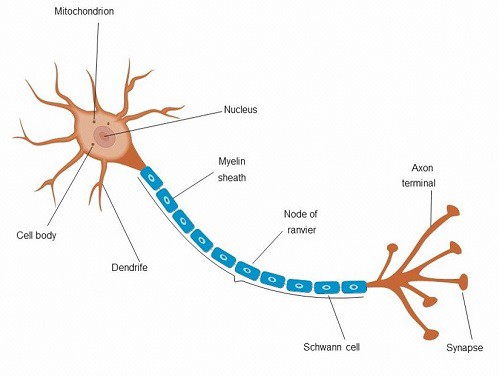
# Example of Figures



Figure D. 1 : American International University-Bangladesh (AIUB)

The Figure [D.1](#_bookmark28) represents beauty of the AIUB campus.

*APPENDIX D. EXAMPLE OF FIGURES*



1. Anatomy of a multipolar neuron (b) Architecture of a artificial neuron

Figure D. 2: Example of placing images side by side

**Appendix E**

# Example of Tables

Here is a really simple table [E.1](#_bookmark31).

Table E. 1: AIUB currently operates under four distinct Faculties x

|  |  |
| --- | --- |
| **Number** | **Name** |
| 1 | Faculty of Science and Technology (FST) |
| 2 | Faculty of Engineering (FE) |
| 3 | Faculty of Business Administration (FBA) |
| 4 | Faculty of Arts and Social Sciences (FASS) |

Here is another example of table row merged [E.2](#_bookmark32).

Table E. 2: Row span example

|  |  |  |
| --- | --- | --- |
| col1 | col2 | col3 |
| Multiple row | cell2  cell5 cell8 | cell3  cell6 cell9 |

Here is another example of controlling table width [E.3](#_bookmark33).

Table E. 3 : Test Table

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| One | Two | Three | Four | Five | Six | Seven | Eight | Nine | Ten | Eleven | Twelve | Thirteen | Fourteen |
| 1*.*111 | 2*.*222 | 3*.*333 | 4*.*444 | 5*.*555 | 6*.*666 | 7*.*777 | 8*.*888 | 9*.*999 | 0*.*000 | 1*.*111 | 2*.*222 | 3*.*333 | 4*.*444 |

**Appendix F**

# Example of algorithm procedure

**Algorithm 1:** Example code

**Input:** A graph *G* **Output:** A vertex of *G* **Data:** Testing set *x*

∞

**1** ∑

*i*=1

:= 0 // this is a comment

/\* Now this is an if...else conditional loop \*/

**2 if** *Condition 1* **then**

**3** Do something // this is another comment

**4 if** *sub-Condition* **then**

**5** Do a lot

**6 else if** *Condition 2* **then**

**7** Do Otherwise

/\* Now this is a for loop \*/

**8 for** *sequence* **do**

**9** loop instructions

**10 else**

**11** Do the rest

/\* Now this is a While loop \*/

**12 while** *Condition* **do**

**13** Do something

Example of writing algorithms is shown here [1](#_bookmark35).

**Appendix G**

# Example of Code

## G.1 Find the greatest number from a list of numbers in *Python*

a=[1,2,3,4,6,7,99,88,999]

max= 0

for i in a:

if i > max:

max=i print(max)

*End quote goes here.*