

Musarat Hussain

# EXPLORING CLUSTERING, DEEP LEARNING, AND LLMs IN TEXT CLASSIFICATION

Faculty of Information Technology and Communication  
Master's thesis  
August 2024

# Abstract

Musarat Hussain: EXPLORING CLUSTERING, DEEP LEARNING, AND LLMs  
IN TEXT CLASSIFICATION

Master's thesis

Tampere University

Master's Degree Education in Computing Sciences

August 2024

---

Natural language processing and sentiment analysis are important in the current era. Many people are working in this domain to understand the human language and try to classify it. Humans are more like to express their opinions using open text rather than pre-defined questions. This study uses different unsupervised, deep learning, and **Large Language Models** to classify text data. Two datasets of different topics have been chosen for training and evaluating all models. It included Apple product reviews and airline tweets. The study aims to evaluate the performance of different classification algorithms and models to see which one is more accurately classifying tweets as compared to others. This study is also important as the comparison is done with the latest model of openai which is GPT-4. The findings of this research demonstrate that among all the algorithm tested, BERT based model and GPT-4 exhibit superior performance. The Roberta-based bert model depicts 81% accuracy on the Apple dataset while the bert-based-uncased model outperformed others on the airline dataset with an impressive accuracy of 95%. GPT-4 also depicts strong results with an accuracy of 79% for the Apple dataset and 85% for airline sentiment. This is a strong indication that future model of openai or other LLM models might surpass the BERT model.

These results and analysis show that LLM models like BERT and GPT-4 are more effective for sentiment classification as compared to traditional machine learning and deep learning algorithms. It is also worth noting that LLM models require less cleaning and pre-processing of datasets as those are already pre-trained models. This feature enhances efficiency and usability. This research provides potential for LLM models in text classification which also offer valuable insights for future research. Overall, this study highlights the power of the LLM model over a conventional model for sentiment data classification. It provides a detail comparison of their performance and to discuss the implifcaiotn of these method for the field of natural language processing.

**Keywords: LLM, Sentiment, Deep learning, Classification.**

The originality of this thesis has been checked using the Turnitin Originality Check service.

# Contents

1	Introduction . . . . .	1
2	Literature Review . . . . .	5
2.1	Classification Using Deep Learning Algorithms . . . . .	5
2.2	Classification using BERT . . . . .	6
2.3	Classification Using Other Large Language Models . . . . .	9
2.4	Recent Developments and Future Directions . . . . .	12
3	Methodology . . . . .	13
3.1	Data Collection . . . . .	13
3.1.1	Apple Sentiments . . . . .	13
3.1.2	Airline Sentiments . . . . .	13
3.2	Pre-Processing . . . . .	13
3.2.1	Data Cleaning . . . . .	14
3.2.2	Tokenization . . . . .	14
3.2.3	Stopword Removal . . . . .	14
3.2.4	Lemmatization . . . . .	14
3.3	TF-IDF vectors . . . . .	15
3.3.1	Term Frequency (TF) . . . . .	15
3.3.2	Inverse Document Frequency (IDF) . . . . .	15
3.3.3	TF-IDF score . . . . .	15
3.4	Model Selection . . . . .	15
3.4.1	K means . . . . .	16
3.4.2	Hierarchical Clustering . . . . .	17
3.4.3	DBSCAN . . . . .	18
3.4.4	Feedforward Neural Network . . . . .	20
3.4.5	Recurrent Neural Network . . . . .	21
3.4.6	BERT . . . . .	21
3.4.7	GPT-4 . . . . .	23
3.5	Tools and Software . . . . .	23
4	Results and Discussion . . . . .	25
4.1	Evaluation Metrics . . . . .	25
4.1.1	Accuracy . . . . .	25
4.1.2	Precision . . . . .	25
4.1.3	Recall . . . . .	25
4.1.4	F1-score . . . . .	26
4.1.5	Confusion Matrix . . . . .	26

4.2	K-mean . . . . .	27
4.3	Hierarchical Clustering . . . . .	27
4.4	DBSCAN Clustering . . . . .	27
4.5	Feedforward Neural Network . . . . .	28
4.5.1	FNN (Apple Sentiment) . . . . .	29
4.5.2	FNN (Airline Sentiment) . . . . .	30
4.6	Recurrent Neural Network . . . . .	30
4.7	BERT . . . . .	33
4.8	GPT -4 . . . . .	36
4.9	Models Comparison . . . . .	40
4.9.1	Intra BERT comparison for apple sentiment . . . . .	40
4.9.2	Intra BERT comparison for Airline sentiment . . . . .	41
4.9.3	All model comparisons for apple sentiment . . . . .	42
4.9.4	All Model Comparison for Airline Sentiment . . . . .	43
4.10	Lessons Learnt from the Experiments . . . . .	43
4.10.1	BERT and GPT-4 superiority . . . . .	43
4.10.2	Tradition Approaches Limitation . . . . .	44
5	Conclusion and Future Work . . . . .	45
6	References . . . . .	47

# 1 Introduction

With the advancement in social media and communication technologies, many people prefer to participate in social networks to express their opinions [17]. The process of identifying and extracting subjective information from human language like text data has taken enormous importance. Classification of such views is a heated topic in natural language processing. It is also the aim of this research that different machine learning, deep learning, and LLM models will be used to classify those tweets.

Text Data classification is the process of assigning predefined labels, or categories automatically to the text data or document based on its content. This task is important for a wide range of applications and domains including topic modeling, spam filtering, sentiment analysis, and document categorization. Effective classification and analysis of text data facilitate information retrieval, enable better decision-making, and unlock valuable insight across multiple domains. Another important application of sentiment analysis is to answer customer queries i.e. is customer service. It is the classification of customer queries, complaints, or feedback and route them to the relevant person, or department, or to help in generating automated responses. Topic modeling means organizing text documents into themes or coherent topics. Document Categorization is the grouping of text data into predefined categories (blog posts, product reviews, articles, and news).

Unsupervised algorithms are a class of machine-learning techniques that find patterns and does features extraction from data without any need for labeling dataset manually. Or any pre-defined target variables. This kind of algorithm aims to uncover the underlying structure and relationship of points within the dataset. It uses several techniques including dimensionality reduction, clustering, and anomaly detection. By learning from the data itself, it can discover hidden connections and novel insights that might not be very obvious to humans.

Furthermore, deep learning models are a subsection of the machine learning domain. It uses artificial neural networks that contain many hidden layers to understand hierarchical and complex data representations. Deep learning is good for understanding the pattern inside unstructured data including text, images, and audio. These models can learn features from data, without doing any manual feature engineering. It is useful for performing well in domains like NLP, speech recognition, and computer vision [3].

The last type of model used in this study is the Large Language Model (LLM). These models are trained on large data which helps them to understand complex

patterns inside the dataset. These models like GPT, and BERT can understand the complexity of human language, question answers, and other tasks.

Traditional approaches to classification and sentiment analysis have relied on rule-based methods or lexicon-based techniques, which involve manually curated dictionaries of words associated with positive or negative sentiment. However, these traditional methods often struggle to capture language's nuanced and contextual nature, leading to suboptimal performance, especially on more complex or domain-specific text [18].

In the recent era, the rapid advancement of deep learning, NLP, and language models has revolutionized the domain of sentiment analysis and processing. The invention of different deep learning and LLM played a major role in increasing the performance of such tasks. Many such (Deep learning) algorithms including Convolutional neural networks (CNNs) and Recurrent neural networks (RNN) excel at learning the hierarchical representation of text. It means that deep learning models can understand complex patterns inside the dataset. It is critical for understanding the sentiment, which helps analyze sentiment data. Deep learning models can learn relevant features from raw text, improving the model's accuracy and efficiency.

Recently one special type of mechanism called self-attention enables the model to give importance to each word depending on the sentence it belongs to. It is called BERT which has enormous results with text data as compared to deep learning models [5]. Three different variants of BERT models are in the scope of this research. Comparison of such model with the latest model of openai like GPT-4 has not been made quite a lot. There are not any proper study on this kind of comparison because of the newest release of GPT-4. Overall, the comparison of deep learning, BERT, and GPT-4 for different use cases is under scope.

Two datasets including Apple product sentiment and US airline sentiments have been taken from freely available sources like dataworld and Kaggle. Apple sentiment dataset contains the sentiment of people in the form of tweets about the Apple product. It shows how people feel about the product overall and how are their reaction. As Twitter is a platform that is accessible to everyone in every part of the world, this data contains opinions from very broad people around the world. It also means that the dataset is very neutral and is very useful for the improvement of the product in the future. The second dataset used is US Airline sentiment data. This dataset is also consisting of tweets about airlines, and their operation from their passengers. Similar to apple sentiment, these are also very neutral and from the passenger who actually use flights to travel. Different models including K-means, DBSCAN, hierarchical clustering, Feedforward neural network, Recurrent neural network, BERT, and GPT-4 have been used to classify tweets into positive,

negative, and neutral categories.

This thesis explores the efficiency of different methodologies for sentiment classification focusing on three different domain models for a specific use case. These domains include machine learning, deep learning, and large language models (LLM). The aim is to provide a comprehensive comparison of all the models used for training and assess their evaluation ability. By comparing different metrics of all models, this study will provide valuable insights for practitioners and researchers in the field of investment classification. It will also highlight how the LLM model might perform better in the future.

Through such kind of study, the aim is to advance the understanding regarding sentiment classification, and understanding of how different models perform when used for classification tasks.

The real motivation behind this topic is the need for text data classification due to the rapid growth of digital data, particularly text data from different sources like online reviews, enterprise documents, and social media. Traditional models require extensive labelling of datasets and it can be time-consuming and expensive for large space text corpora. [17,18]. In recent years, the advancement of deep learning models, and large language models including BERT and GPT-4 has enormous benefits and ease of classifying text data[19,20]. Large language models don't need more data to be trained first and then start classifying, which makes LLM very interesting, it also generates significant interest in its potential for text data classification tasks.

My thesis aims to comprehensively analyse and compare unsupervised, deep learning, and large language models for text data classification. It will especially address the following research question.

- a. How do various Classification techniques perform and compare in terms of accuracy, precision, recall, and F1-score for sentiment classification tasks on a Twitter dataset?
- b. Can the GPT-4 model, with its impressive text data understanding and generation capabilities be effectively used for text data classification, and how does its performance compare to other BERT and other deep learning models?

By answering all these questions and comparing the results of all the models and techniques used for text data classification, this thesis will contribute to the understanding of comparing different text data classification methods, enabling students, researchers, and practitioners to make informed decisions when selecting the most approaches approach and model for doing text data classification. This will open

up the discussion of comparing LLM models with each other and will continue until the best one is found and started using it.

To answer all these research questions different clustering techniques like K-mean, Hierarchical, and DBSCAN clustering are used to classify tweets into positive, negative and neutral tweets. Silhouette score is used for evaluation. Furthermore, from the deep learning domain, feedforward neural networks, and Recurrent neural networks have been used to classify tweets. In the end from the LLM domain, BERT and GPT-4 models have been used to classify tweets. Their evaluation metrics have been used to evaluate and compare all the models.

The upcoming sections of this thesis are a literature review where relevant studies have been mentioned, methodology including what all models are and their architecture, results and discussion, content a detailed discussion about the results achieved, and a conclusion that concludes the whole thesis.

## 2 Literature Review

Initially sentiment analysis and classification were relied on the traditional machine learning models. Despite the simplicity of machine learning models, it laid the foundation of more complex and sophisticated models. With time the nature and size of dataset is changed. As the data increases in size multiple approaches have been used to classify sentiments into their respective classes. These domains included machine learning, deep learning, and large language models. Some deep learning approaches including CNN and LSTM perform quite well in more complex data as compared to traditional machine learning models. Deep learning is a dominant method of NLP. Many papers have been published related to text classification tasks by using various deep learning models including RNN, FNN, etc. BERT has been introduced as a groundbreaking model that with the help of bidirectional training to capture full context of the sentence leading to better prediction and performance. This advancement is signification in natural language processing. The ability of the BERT to understand the context of the sentence make it helps to achieve state of the art performance across various NLP task cinlusing sentiment analysis. This ability of BERT make it a highly affective tool for determine sentiment accurately in diverse task [4]. It offers great improvement to learning embedding from Scratch. The author of [16] introduce GloVe (Global Vectors for Word Representation in 2014 which has become a basic technique in natural language processing. GloVe is a new method that is used for getting vector representation of words. The idea on which the model is based is that it should be able gather global statistical information. It is slightly different from traditional word embedding technique as those rely on local context. GloVe has a special way of generating word vector by incorporating world co-occurrence matrix. GloVe vectors capture meaningful semantics. Using GloVe model shows better performance. GloVe has the ability to balance the trade off between performance and computational efficiency. It can train on large dataset relatively quicker as compare to deep learning models. The model's semantic relationships and encoding syntactic has made it a widely used tool in the field and it is contributing in the advancement of NLP task and applications.

### 2.1 Classification Using Deep Learning Algorithms

In a study [3], the author discusses the advancement of deep learning techniques over traditional algorithms of machine learning. The author did a comprehensive analysis and application of tasks like sentiment analysis, and text data classification using RNN, CNN, and transformer model. The study found the superior performance of

all these models in understanding the complex contextual information, and complex patterns inside the dataset. Although the author faced many challenges with the need for a large, labeled dataset, and computational complexity, still the potential for deep learning models having better accuracy and efficiency of text data analysis is proof of ongoing innovation in this field.

In 2019, another study proposed a deep neural network with a sentiment attention mechanism for text sentiment classification, underscoring the significance of attention mechanisms in improving sentiment analysis tasks [9]. Basically, this study combines the sentiment attention mechanism and deep neural network for text sentiment classification. This approach increases the ability of the model to focus on specific words and phrases of sentiment. It improves the sentiment classification accuracy. By using both attention mechanism and neural network, this method achieves supervisor performance as compared to normal neural network models. This method is also versatile and could be apply to diverse text data which makes is a good contribution to sentiment analysis field.

Moreover, Wang et al in 2021 have explored deep learning structures for sentiment classification, emphasizing the importance of optimizing models like RNN and CNN to enhance text sentiment classification performance. The approach of sentiment classification that uses weak tagging information for model accuracy improvement is explored. For this purpose, BERT base model and structure are integrated with a weak supervision mechanism that took large-scale weakly labeled data for training. Furthermore, it fine-tunes the model with accurately labeled data. This way the labelling cost is reduced, and accuracy is increased as compared to other traditional machine learning models. Weak tagging information in this model architecture helps the model to capture diverse sentiments in text data [13]. Another similar study has been conducted [7] where the author proposes a framework of deep learning for cross-domain classification. This approach considers weight adjustment mechanisms and improves cross-entropy loss. This proposes approach consider different domains of sentiment and addresses the challenge of shifting sentiment knowledge. It enhances the model generalization. Propose method also takes BERT for taking contextual understanding, perform very well than baseline model in cross domain sentiment tasks. This method is a valuable contribution int he field of sentiment data analysis domain due to its speciality in handling domain variations [7].

## **2.2 Classification using BERT**

Deep learning and large language model's accuracy have been compared for a long time. In 2024, comprehensive research has been conducted on the BERT model application in the domain of sentiment analysis and classification. It demonstrates the

supervisor performance of the BERT model over deep learning and other traditional models. This study also throw light on the supervisor ability of BERT model in capturing contextual nuances through its special encoding mechanism. This bidirectional encoding mechanism is leading it to better accuracy, and robustness in text data classification task. Moreover, several factors including pruning, computational efficiency, and fine tuning make BERT more powerful tools for sentiment analysis. Overall, this study suggest that BERT enhances the classification understanding and prediction and consider it as a critical advancement int the field of natural language processing [24].

Sentiment classification is a fundamental task in natural language processing, and the utilization of deep learning models like BERT has demonstrated significant progress in this field. Several studies have investigated the application of BERT in sentiment classification tasks. For example, as discussed above that BERT has been introduced as a powerful model that can be fine-tuned for various tasks with minimal modifications, showcasing its effectiveness in tasks like question answering and language inference. they also focused on fine-grained sentiment classification using BERT, highlighting its capabilities in solving intricate sentiment analysis tasks. Additionally [6] proposed a sentiment classification model that combines BERT with an adaptive sentiment dictionary, demonstrating the potential of leveraging BERT in sentiment analysis. It is a novel model with capability of BERT and adoptive sentiment dictionary to enhance the accuracy of sentiment analysis. Firstly, the BERT is used for feature extraction due to its contextual understanding, then the adoptive sentiment dictionary refines the classification of sentiment by using sentiment-specific information. This hybrid approach help in achieving stat of the art result while classifying dataset into different classes. The integration of new method (adoptive sentiment dictionary) addresses and overcome the limitation of the BERT model standalone, making this approach a significant advancement in the field of natural language processing [6].

In the realm of sentiment classification in specific domains, the author of [25] focused on sentiment classification in online health communities, highlighting the importance of additional pre-training tasks to enhance sentiment classification accuracy for complex texts. Two approaches including LDA for topic modelling and BERT for contextual understanding are combined to identity the information users seek. This approach improves the satisfaction and engagement by finding the important content to meet user needs. It also wider the usage of BERT model in various text data analysis task where to understand the user needs is crucial [25].

Additionally [15] and [22] explored sentiment classification in the context of COVID-19 discussions and social media posts, respectively, showcasing the versatility of

BERT in analysing sentiments across different domains. [22] focuses on analysing negative sentiments in China reflecting the negative opinion spreading by people during early pandemic situation. The aim of the study is to understand the emotion of the public regarding the pandemic situation and to understand the capability of the BERT model to sentiment data analysis. This study also makes it clear that to capture negative sentiments from the dataset, BERT is a highly affective way of doing it. Its special architecture helps in understanding the context and language. This ability is making it more superior than traditional models. This study also reveals the negative trends during different phases of pandemic. This kind of information is also crucial for policymaking and health officials. This study highlighted the application of BERT in real time monitoring of sentiment of the people. This ability is important for timely interventions and decision making. On the other hand, the author of [15] introduces redBERT model combine BERT model with topic discovery technique to analysis social media discussion. The focus of this study was to use BERT's deep learning ability and analyse huge amount of COVID-19 pandemic data. The finding of this research is that roberta model demonstrate high accuracy in classification of sentiment. This classification helps to understand public opinion regarding the on-going pandemic. This kind of study is also good for government officials in decision making. Also, the takeaway from this study is that roberta base model perform better than the traditional machine learning models in both domain including sentiment classification and topic discovery. In short both these studies focuses is to explore BERT's effectiveness in classifying sentiments and to provide important insights in public opinion, trend during COVID-19 pandemic. This kind of studies specially during situation like pandemics are crucial for the government officials in decision making.

Michael in 2024 [12] explore the classification of tweets related to Ghanaian football during 2022 FIFA world cup. By applying sentiment classifying techniques the study successfully classifies the tweets in to negative, positive and neutral tweets. These finding highlights the opinion of fans and explore the fluctuation in the sentiments of the people in response to different matches and events. This study also illustrates the application of sentiment analysis in understanding of public opinion in the context of sports events. It provides important information to stake holders and sports managers. Author of [1] proposes a hybrid technique combining deep learning with fine-tuned BERT models for domain-specific sentiment analysis, showcasing the adaptability of BERT for specialized tasks. By combining multiple fine-tuned models of BERT, the study increases the accuracy and robustness of sentiment classification. The dataset used in this study was curated from software engineering discussion, demonstrated the effective of the proposed approach in considering domain specific nuances. This fuzzy technique of combining outperforms

the traditional techniques used for sentiment classification.

This paper [11] introduces RoBERTa, a robustly optimized version of the BERT language model. The study evaluates its performance on various text classification tasks compared to BERT and other deep learning models. The key improvement of this RoBERTa base model was that it trained on a larger dataset, used a diverse corpus of datasets while training, for a longer duration, and took more optimized steps in training as compared to another variant of the BERT model. It also employed a dynamic masking strategy which make it more special and improve performance. The token of the mask is randomly picked at each training step rather than selecting the same mask for each step. The study also illustrates that RoBERTa is performing better than the simple BERT model. Compariosn of RoBERTa with another variant of BERT like bert-base uncased and bert-large uncased is studied in this thesis as well where RoBERTa shows some promising results when compared with others. This model's improved contextual understanding and robustness make it particularly effective for sentiment analysis tasks. It allows a more accurate interpretation of sentiment in diverse domains of sentiment analysis.

### **2.3 Classification Using Other Large Language Models**

Author of [18] introduces GPT-2 as a large language model that excels in unsupervised multitask learning. It also shows the parameter used for GPT-2 to train which is 1.5 billion. In this seminal work that introduced GPT-2, the authors found that GPT-3.5 (a later version of GPT-2) exhibited strong performance on various text classification tasks, and often performed well. It also performs well on task like summarization, translation and many more without need of nay large data for training or understanding purpose. It has been trained on billions of words of text data from different books, the internet, and other sources. It is designed to leverage the power of self-attention mechanisms and transformer architecture to learn, understand, and capture the complex structure, and pattern of the human-generated text.

The author of [20] discusses the capabilities of large-scale language models, such as GPT-2, to perform a wide range of natural language processing tasks, including text classification, without the need for task-specific supervised training. In this study, the author trained GPT-2 on diverse corpus of internet text data in an unsupervised manner. Then the performance of the models is evaluated on certain tasks. The study also found that the GPT-2 is performing better without any fine-tuning. This study also explores the scaling effect of the model size and training data. It found that as the training data size increased, the performance also increased. GPT-2 model serves as a powerful and versatile text data processing tool. Large language models like GPT series are pre-trained which means it doesn't need any further

training or more training data for better performance. It just needs a few prompts to understand what it needs to do and then start doing these natural language processing tasks.

GPT-3 is presented as a large-scale autoregressive language model that can be effectively fine-tuned for various text classification tasks with minimal labelled data [2]. The author of the paper introduces GPT-3 as a groundbreaking language model that perform well in few shots learning. It does not need large training like other models. GPT-3 is trained on 175 billion parameters, gives it extra ability to perform well while doing natural language processing tasks with minimum training data available. It is found in the same study that with less data, GPT-3 often outperforms BERT on the text classification task. This shows how effectively GPT-3 can perform classification tasks when there is minimal data available. It requires minimal fine-tuning and can be adapted to new tasks. This study also depicts that BERT generally performs better in standard supervised learning setups with large datasets while GPT-3 is better when the data is not enough. The model's adaptability to multiple task and performance across several benchmark shows it ability and potential for application in natural language processing and other areas tasks. This models ability also highlighted how LLM advances and how it is impacted on flexible and general-purpose AI systems [2].

This study [26] introduces Microsoft 2017 speech recognition system, illustrating the latest advancements in neural network training technique and architecture. This system achieve excellent result in speech recognition, reduces word error rate. The performance comparison of a deep learning-based speech recognition system with a GPT-based language model is done and demonstrates the superior performance of the GPT-based model on certain text classification tasks. The comparison of GPT and deep learning models is not only done on one specific domain but also on sentiment analysis, and topic classification. It also found that GPT has supervisor language understanding as compared to the deep learning model. It explores the strength of both deep learning and the GPT model. It uses GPT for speech recognition which highlight a strength of GPT in that domain.

Another study on the Massive Multitask Language Understanding (MMLU) benchmark provides a comprehensive evaluation of BERT, GPT, and other deep learning models, further demonstrating the impressive performance of large language models, particularly GPT-based models, on a broad range of natural language understanding tasks [8]. This study also highlighted the impressive accuracy of large language models. It include GPT based on a wide NLP tasks. The author also suggests that continuous growth, development, scaling, and improvement of large language models with the ability to understand human-like text can lead to significant advancement

and precision in machine learning field understanding. In the future, these models might perform very well even with a few shots and prompts of the data for hints to the model.

Another study [19] compares the performance of one large language model (Gopher) to other prominent LLM models including T5 (from Google) GPT-3 (from Openai) BERT (from Google) This study focuses more on the scaling of language models, particularly the growth, and development of the Gopher model created by Deepmind. It is a large-scale language model from Deepmind. It is also a transformer-based model that is trained on diverse form of data including text data, books, web pages, and different scientific literature. One purpose of this model is to explore the capability of a large language model as it grows in size, scale, and complexity. For comparison of Gopher models to other large language models, wide range of tasks have been chosen. It includes Commonsense reasoning, symbolic mathematics, natural language inference, and reading comprehension. This study uses F1 score, and perplexity as its evaluation metrics to assess the model. The evaluation shows that Gopher outperformed T5, GPT-3, and BERT in majority of evaluation metrics [19]. Gopher exhibited strong few-shot or zero-shot learning. It depicts it captures new data patterns with a very minimal dataset. This study also illustrates that continued scaling can lead to better performance.

The author of [14] did comprehensive comparison of GPT with many traditional language models. The performance of all models has been evaluated against the traditional machine learning models in performing sentiment analysis. The study depicts that large language models of openai like GPT 3.5 and other similar models is significantly performing well as compared to traditional machine learning models. These traditional models include Random Forest, SVM, and Naive Bayes, and many more in terms of accuracy. The ability of large language models to understand the pattern in new data is contributed to its excellent performance. Also, one important strength of GPT base models is to handle the context of the sentiment and perform accordingly which is not the case in traditional models. Traditional models of machine learning, and deep learning usually struggle with context understanding whereas GPT uses its pre-trained features to understand the context very well. GPT is also better for its less pre-creation needed. Other models require extensive feature engineering before training models for improving performance which is not the case with GPT. This study is also evidence of GPT's superior scalability and adaptability. It performs well across different domains without the need for retraining extensively. Traditional models need separate training each time when any new task is going to be performed with them. In short, this paper represents GPT models as the advancement in the field of sentiment analysis.

## 2.4 Recent Developments and Future Directions

Recent research in this field is continuing to build on success of BERT, exploring different variant of it like RoBERTa, and large language models like GPT-3 and 3.5. It tries to achieve more robust and better performance from different models. Furthermore, different studies which are comparing GPT-3 with traditional approaches and models including deep learning for classification of sentiment have shown the increasing potential of large language models int the field of sentiment analysis.

GPT-4 is a very latest and updated model of Openai. There is not any paper been published about the comparison of performance between GPT-4 with BERT or deep learning models. This is the specialty of this study and the novelty that it is comparing the performance of GPT-4 with BERT and deep learning models to know the difference and predict the ability of future LLM models.

From a different perspective, INCEPT [27] utilizes multiple representation methods to detect duplicate text pairs. INCEPT involves using a stacking ensemble of pairwise vector distance measurements that are computed from multiple text representation methods. A stacking classifier then utilizes these distance scores as input and learns to identify duplicate posts.

## 3 Methodology

In this section, complete steps and procedure that is followed to conduct research have been discussed. It provides a detailed description of all the models, their architecture, and relevant terms. It encompasses the process of data collection, cleaning, pre-processing, model implementation, training, and evaluation. This section aims to provide a clear replicable path for achieving the research objective.

### 3.1 Data Collection

This research project aims to classify Twitter data into clusters using different models. For this purpose, two different datasets have been taken to compare models better. The apple sentiment dataset is taken from data.world.<sup>1</sup>

2. The airline sentiment dataset is taken from Kaggle.<sup>2</sup>

#### 3.1.1 Apple Sentiments

This dataset contains tweets about Apple products and the size of the dataset is about 3.8 thousand tweets about the product. It contains all those tweets that contain @apple, or AAPL. The sentiments of the tweets are labeled as positive, negative, and neutral.

#### 3.1.2 Airline Sentiments

These datasets contain tweets that show how travelers reacted about airlines in February 2015 on Twitter. The size of the dataset is about 14 thousand tweets. The sentiments of the tweets are positive, negative, and neutral. This dataset contains tweets about major US airlines from February 2015. These tweets are labeled as positive, negative, and neutral tweets.

### 3.2 Pre-Processing

Pre-processing is a crucial step to be taken for preparing the raw data for analysis and modelling. Proper steps of pre-processing can enhance the quality of the dataset which makes it more suitable for training different deep learning and machine learning models. Following are the steps taken to meet the aim of pre-processing.

---

<sup>1</sup><https://data.world/crowdfunder/apple-twitter-sentiment>

<sup>2</sup><https://www.kaggle.com/datasets/crowdfunder/twitter-airline-sentiment>

### 3.2.1 Data Cleaning

Cleaning the dataset is a critical step before taking data for training in any model. It ensures that the dataset is now prepared and ready to move to the next steps. Following are the steps taken to clean the text for the model.

- a. All missing values are replaced with empty string to avoid any error.
- b. All text is converted to lowercase to adopt uniformity
- c. Remove all urls from the text
- d. Remove all mentions from the tweet data
- e. Remove all hashtags, extra white spaces, and special characters

### 3.2.2 Tokenization

Tokenization is the breakdown of text into individual tokens or words. It is quite important pre-processing for deep learning models as it converts pure text to tokens. These individual words are easy to process and use in model training. For tokenization, the toolkit(nltk) library is used. Furthermore, punkt tokenizer models were downloaded to ensure that the library can tokenize the text.

### 3.2.3 Stopword Removal

Stopword removal removes less meaningful common words to improve the classifier's performance. Examples of such words are article, conjunction, and preposition. Such words create an issue while converting words to TF-IDF vectors as these words have no importance. Nltk library is utilized for this, and the resources are downloaded to ensure the library can access the list of stop words.

### 3.2.4 Lemmatization

It is the process of reducing words to their root or base form called lemma. This step ensures that different forms are treated as a single item, which helps normalize the text. Lemam is a dictionary form of a word.

WordNetLemmatizer from nltk is used for the lemmatization of text. Lemmatization helps in

Normalization

Improve accuracy

Semantic meaning

### 3.3 TF-IDF vectors

TF-IDF stands for Term Frequency -Inverse Document Frequency and it is a statistical measure which is used to evaluate the importance of each word in the form of digits. It helps in transforming text data to a numerical value which is then used in machine learning models.

There are three components of TF-IDF vectors.

Term Frequency (TF)

Inverse Document Frequency (IDF)

TF-IDF score

#### 3.3.1 Term Frequency (TF)

TF measures how frequently a term occurs in a document. The formula to calculate TF is equation 3.1

$$\text{TF} = \frac{\text{Count of term } t \text{ appear in a document}}{\text{Total number of terms}} \quad (3.1)$$

TF provides the frequency of the term in a specific document.

#### 3.3.2 Inverse Document Frequency (IDF)

IDF is more concerned with the term's importance. IDF measures the global importance of the term throughout the documents. The formula to calculate IDF is equation 3.2.

$$\text{IDF} = \frac{\text{Total number of documents } N}{\text{Number of documents containing term } t} \quad (3.2)$$

#### 3.3.3 TF-IDF score

This score is the product of TF and IDF values.

$$\text{TF-IDF} = \text{TF} \cdot \text{IDF} \quad (3.3)$$

### 3.4 Model Selection

In this section, all selected models for this study are discussed and explained in detail. Following are the different unsupervised and supervised models that have been selected

K-mean

Hierarchical clustering

DBSCAN

FNN

RNN

BERT

GPT-4

### **Reason of Choosing these methods:**

The main aim of choosing a variety of methods, from traditional algorithms of machine learning to more advanced deep learning algorithms and also large language models (LLMs) like BERT and GPT-4, allows for a comprehensive evaluation of text data classification.

**FNN** is simpler and faster to train. No complexity, It provides a good baseline performance for text data tasks. Can be extended by adding extra layers.

**RNN** is Best for sequential data, and we have Twitter data. That is why it is chosen. also is chosen as it can carry information from the previous step. Flexibility to input data for classification.

**BERT** is a state-of-the-art model in the field of Natural Language Processing (NLP). that has revolutionized various text classification tasks. BERT is designed to understand the context of a word based on both its preceding and succeeding words.

Furthermore, **GPT-4** which is a new model of the Openai series is taken to see how this new model is performing as compare to all other existing models.

### **3.4.1 K means**

K-mean clustering is an unsupervised machine learning algorithm which is used for grouping similar data points to the same class. Number of classes is defined before classification. The aim is to split all data points into k number of classes considering the similarity of all data points. Following are the steps involved in performing k mean clustering.

**Choosing the number of clusters k:** The value k shows how many clusters the data should be split on.

**Initialize cluster centroids:** To find the center point so that similar data points could be assigned the same label.

**Assign data points to cluster:** Assign all similar points to one culture.

**Update cluster centroid:** after assigning, update the center point.

Keep assigning and updating the cluster centroid so that all points have been split.

All these models have been used to classify the tweets into different classes. Their performances and metrics have been compared to find the one with better accuracy.

### 3.4.2 Hierarchical Clustering

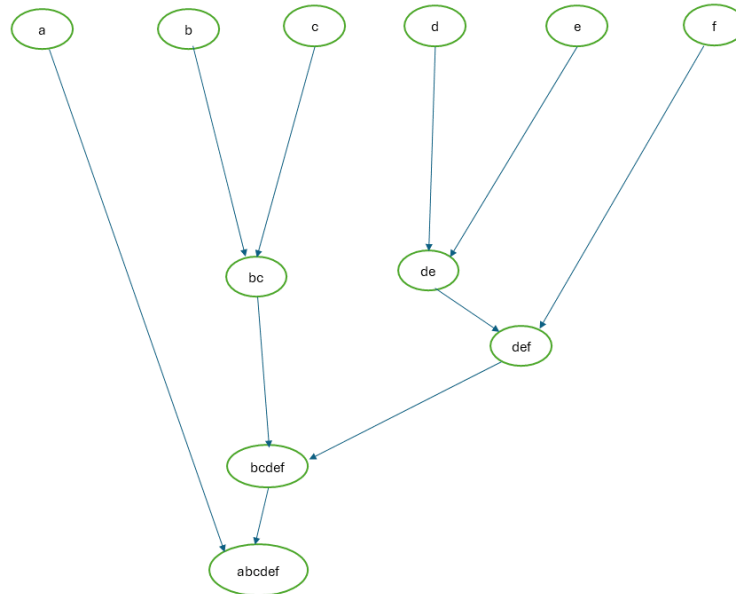
Hierarchical clustering is also an unsupervised machine learning algorithm that groups similar data points into a hierarchy of clusters. It is slightly different from the k mean as hierarchical clustering does not define the number of clusters in advance. Rather it builds a hierarchy of clusters as a dendrogram which helps the user to choose the cluster. Following are the steps to be taken for this kind of cluster.

**Calculating distance matrix:** Hierarchical clustering involves calculating the distance matrix as the first step.

**Merge closest pairs:** It merges the pair which is close to each other to assign the same labels to them.

**Update distance matrix:** After merging the closest pair, the next step is to update the distance matrix to know what other points come closer to each other for merging.

**Repeating merging and updating matrix:** Repeating the processes to get the classified data. figure 3.1.



**Figure 3.1** *Heirarchical Clustering algorithm*

In figure 3.1, it can be seen how hierarchical clustering works. After the first row, it yields clusters a, bc, de, and f. Cutting after the third-row yields bdef, def. In the end, all gather to form cluster abcdef. This method builds a hierarchy from top to bottom by progressively adding or merging clusters. So it is known as hierarchical clustering.

### 3.4.3 DBSCAN

DBSCAN stands for Density-Based Spatial Clustering of application with noise. It is also an unsupervised machine-learning algorithm that doesn't require any pre-defined number of classes. The key feature of the DBSCAN algorithm is that it is a

Density-based cluster

Ability to Detect-Arbitrary-Shaped Cluster

Robust to noise

Furthermore, following the main steps that need to be taken for this algorithm.

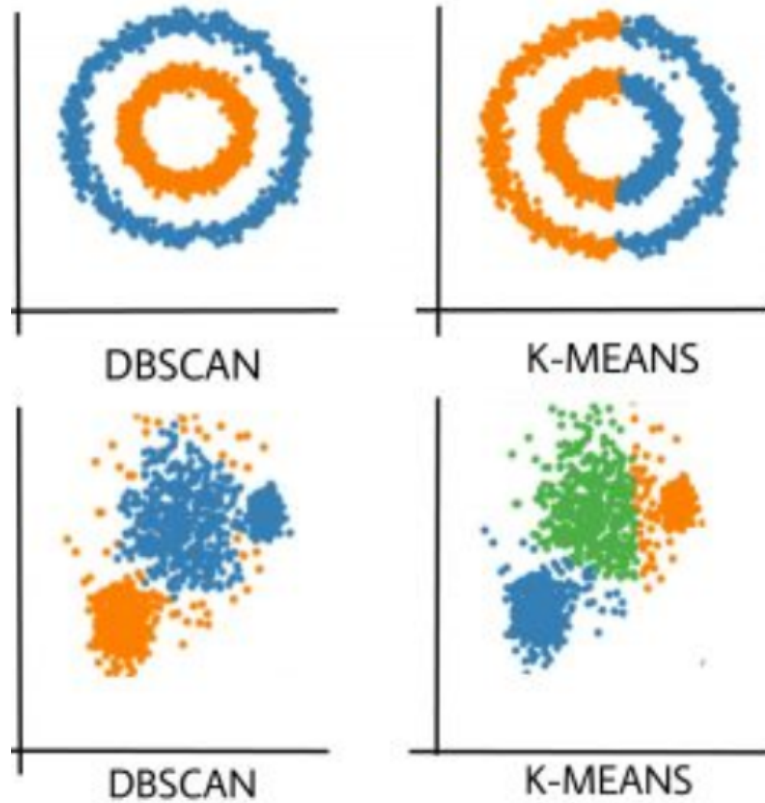
Define Parameter

Identify core points

Expand cluster

Identify Noise

Figure 3.2 illustrates how DBSCAN works differently from the traditional k-mean clustering.



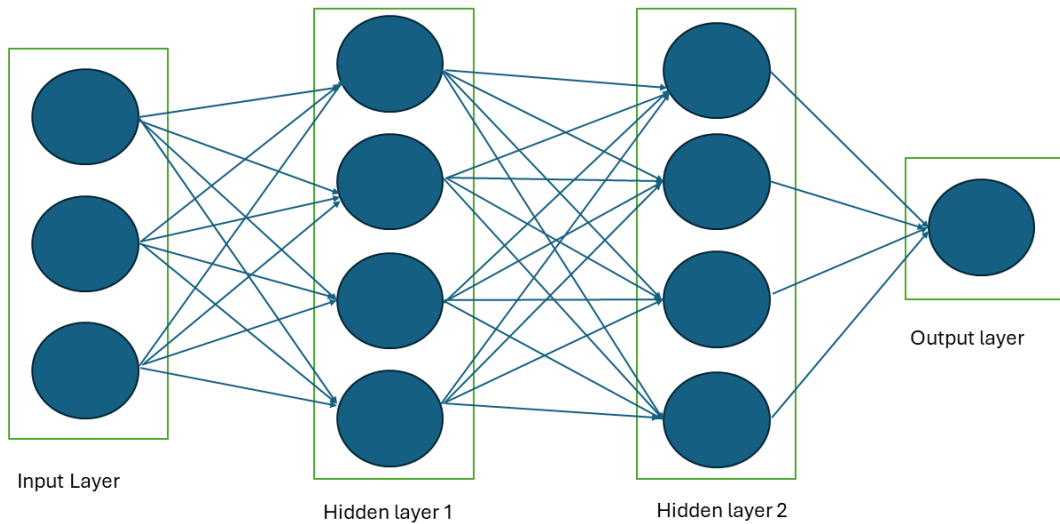
*Figure 3.2 Clusters form in K-mean and DBSCAN*

Figure 3.2 depicts that k-mean solely works on the principle that points are closer to each other and assign all those to the same cluster. However, this is not the case with DBSCAN. It can detect arbitrary shared clusters as in the above figure above.

### 3.4.4 Feedforward Neural Network

FNN is a type of artificial neural network where the connection between nodes doesn't form a cycle. FNN is also known as a multilayer perceptron. It is the simplest AI neural network and serve as the foundation of another complex network [9].

Figure 3.3 shows the architecture of the FNN model. This FNN (also known as multi-layer perceptron) model consists of an input layer, two hidden layers, each followed by a dropout layer to prevent overfitting and an output layer for classification. The ReLU activation function is used in the hidden layers to introduce non-linearity, and softmax activation is used in the output layer for multi-class classification. Due to the presence of multiple hidden layers, it is called a deep feedforward neural network.



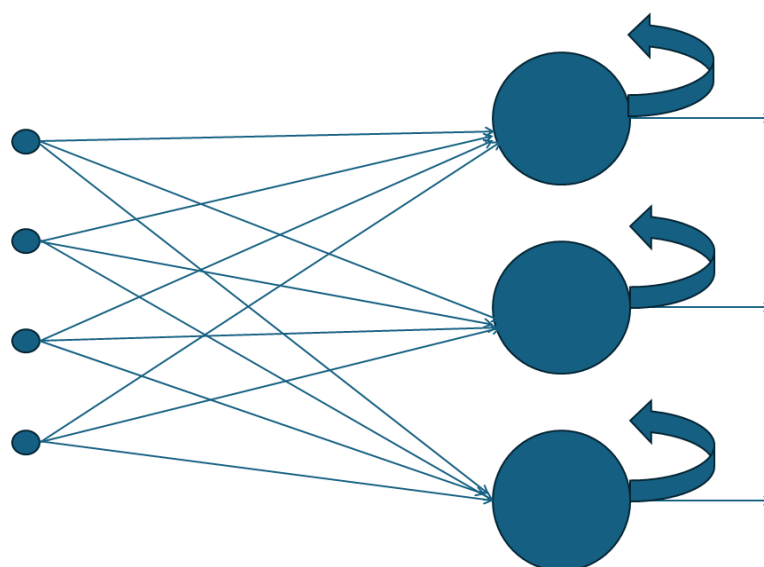
*Figure 3.3 Feedforward Neural Network architecture*

Figure 3.3 depicts the architecture of the FNN model where the first layer is the input layer. The input layer receives the raw input data which consists of features extracted from the dataset. Each neuron in the input layer will represent an individual neuron. There are two hidden layers after the input layer and perform a weighted sum of inputs followed by an activation function to introduce nonlinearity. The last layer in the FNN model is the output layer and it produces the predicted output based on learned features from the previous layers (hidden).

### 3.4.5 Recurrent Neural Network

RNN is a type of artificial neural network that is designed for processing sequences of data. Unlike FNN, RNN has connections that form a directed cycle, which allows information to persist. RNN has a recurrent connection where the output of one step is fed as input to another step [3].

The model that is used is a basic RNN architecture with an embedding layer which is followed by LSTM, and a dense output layer. The summary of the model shows the architecture of the neural network model including types of layers used, number of parameters used, and output shapes. The model used here is sequential.



*Figure 3.4 Recurrent Neural Network architecture*

Figure 3.4 depicts the architecture of the RNN model. It also contains an input layer as its first layer, and it is responsible for receiving a sequence of data. Each input at this point is represented by a vector. There are four inputs, and each of them is responsible for different aspects of the input data sequence. Then there is a hidden layer which processes the data from the input layer sequentially. The hidden layers either pass to output or give it back to the hidden layer for further complex trends in data. The last layer in RNN architecture is the output, which generates the final prediction based on the hidden state of the last layer.

### 3.4.6 BERT

BERT stands for bidirectional encoder representation from the transformer. It is a variant of the Transformer architecture which relies on a self-attention mechanism.

This mechanism lets the model to access the importance of each word in a sentence while processing it. It also helps the model to capture long-range dependencies and relationships which makes it better than traditional deep learning models like FNN, RNN, and CNN [4].

The key components of the BERT model are the following

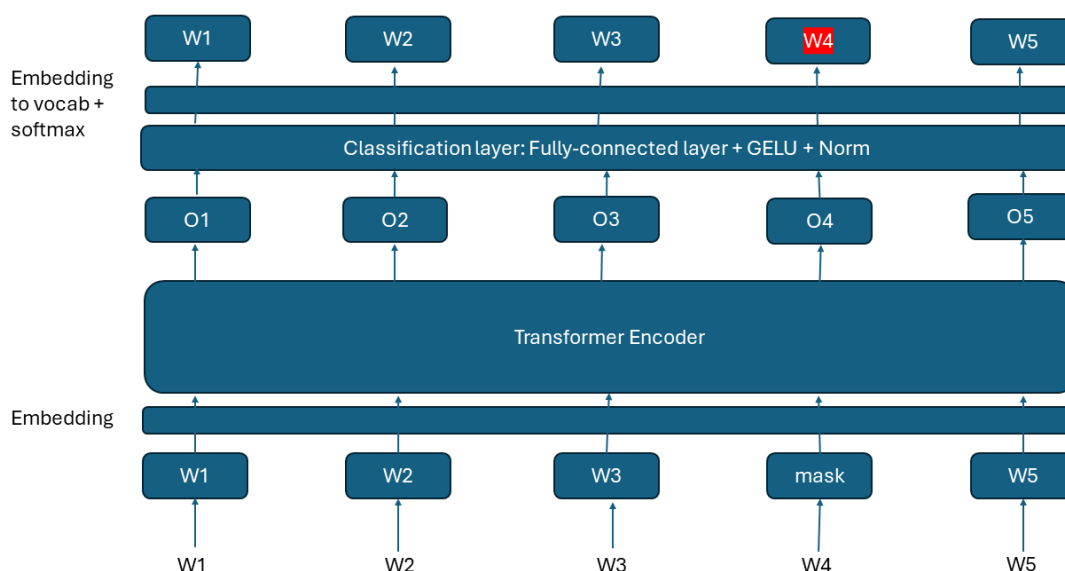
Input representation.

Transformer encoder layer

Pre-trained Objective

Fine-Tuning

Figure 3.5 shows BERT architecture



**Figure 3.5** BERT architecture

Three different variants of BERT will be trained for the classification task. W1, W2, W3, [MASK], W4, W5 are the tokenized input words (W) converted into their corresponding embeddings as seen in the figure 3.5 .

The core of the BERT model is the transformer encoders block which consists of multiple layers of bidirectional transformer encoders. Each encoder layer applies self-attention mechanisms and a feedforward neural network. O1 to O5 represents the output vector from the encoder.

The next layer is classification layer which is a fully connected layer + GELU+Norm. This layer takes the output of the encoder and maps it to the desired output dimension. W1 to W5 are the transformed output after the classification layer.

The last layer in the Bert architecture is the output layer. It maps output from the classification layer to vocab size and applies the softmax function to get probability distribution over the vocabulary. The highest probability indicates the predicted class.

**Bert-Based-uncased** This version is a based version of BERT model which contain 12 layers, 768 hidden units, and 12 attention heads. Uncased means that this version of BERT can't distinguish between lowercase and uppercase letters.

**Bert-large-uncased** This variant is the larger one among all BERT variants with 24 layers, 16 attention heads, and 1024 hidden units. Regarding the case sensitives, it is like BERT-based uncased. It means that Apple and Apple are the same for both variants.

**Roberta-based** This variant of BERT is an optimized version among all which is trained on more data and longer sequences. Architectural point of view, it is similar to BERT-Base. The Roberta-based model has a special ability to change the masking pattern dynamically to improve performance [10].

All three variant of BERT model is using in multiple natural language processing task. These tasks include sentiment analysis, text classification, and much more.

### 3.4.7 GPT-4

GPT-4 is the latest large language model of Openai series which is trained on around 1.7 trillion parameters. It is making it better than GPT 2 and GPT3. For this classification task, an API of GPT-4 is purchased and used. GPT is a large language model that accepts natural text and it doesn't require much cleaning tokenization, or preparation of text before use in the model. Large language models like this are also better if a less-label dataset and just a few shots of data are available. These are specialized few-shots learners which means it can work if few samples of data are there. Unlike deep learning and machine learning, it doesn't need any proper training which makes it a more suitable pick for classification tasks. In this study, a few shots of tweets with labels are given to the model to let the model know what the data looks like and then start giving unlabelled data for classification. API is used for accessing models and classifying data [18].

## 3.5 Tools and Software

The tools and software used for this model training are

**Development Platform:** Google Colab with integrated System RAM and GPU, Jupyter notebook **Computer:** HP

**Openai API** for accessing the GPT 4 model for the sentiment classification task.

**Operating System:** window 10

**Python programming language:** 3.10

**RAM:** 13 Gb

**GPU:** NVIDIA Tesla T4 GPU

**Packages:** Openai, Tensorflow, Seaborn, Sklearn, Matplotlib, nltk

This is the whole experiment setup. The pro version of google colab is purchased for accessing compute units for GPU for faster processing. Without GPU, the model training was taking more time. Also, the reason of choosing google colab is that it is online and there is no need to install any libraries separately as compared to the local system.

## 4 Results and Discussion

For the unsupervised machine learning model, the dataset is tried to classify into 3 different classes. Those classes include positive, negative, and neutral.

### 4.1 Evaluation Metrics

For evaluating models' different evaluation metrics have been chosen. These will be access and compare all models with one another. Following are the metrics that have been used for evaluation.

Accuracy

Precision

Recall

F-1 Score

Confusion matrix

#### 4.1.1 Accuracy

The accuracy of any model is the ratio of correctly predicted instances to the total number of instances.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (4.1)$$

Accuracy is used for model evaluation as it provides an intuitive and simple measure of overall model performance.

#### 4.1.2 Precision

The precision of the model is the ratio of correctly predicted positive labels to the total total positive labels.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (4.2)$$

Precision measure the accuracy of the positive predictions, providing insight into the model's reliability when it predicts a positive class.

#### 4.1.3 Recall

It is also known as true positive rate or sensitivity. It measures the ability of the model to capture positive instances. It is calculated as

$$\text{Recall} = \frac{TP}{TP + FN} \quad (4.3)$$

The reason of choosing recall is used for evaluating and comparing the performance of different models as it shows how well the model is detecting the positive cases,

#### 4.1.4 F1-score

It is the harmonic means of precision and recall.

$$\text{F1 score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4.4)$$

A single measure that balances between precision and recall. It is useful when both false negatives and false positives are considered.

#### 4.1.5 Confusion Matrix

It is a table that is used to describe the performance of the classification model on a set of test data for which true value is already known. It shows the count of true positive (TP), true negative (TN), false positive (FP), and false negative (FN). For the three-class model, this matrix is shown in table ??.

*Table 4.1 Confusion Matrix*

	<b>Predicted Positive</b>	<b>Predicted Neutral</b>	<b>Predicted Negative</b>
Actual Positive	TP	FN	FN
Actual neutral	FT	TP	FN
Actual Negative	FP	FN	TP

**TP** Stands for true positive it shows the number of instances correctly predicted as the actual class.

**FN** Stand for False negative. It depicts the number of instances of the actual class incorrectly predicted as another class.

**FP** Stands for false positive. It is the number of instances of other classes incorrectly predicted as the actual class.

It is used to shows the model's overall performance as it provides a detailed breakdown of model performance across different classes. It allows the researcher to see where the model is performing badly and making errors.

## 4.2 K-mean

This is the first algorithm that has been used in this study to classify the tweets into different clusters. The goal of clustering is to identify the natural grouping inside the dataset and assign the same label to them. The metric that is used to evaluate the model performance is the silhouette score.

Silhouette score is defined as the similarity of an object to its clusters compared to the other cluster. The range of this score is -1 to +1 with values close to 1 representing better defined clusters. Following are the results achieved as a result of k-mean clustering.

All the tweets have been classified into different clusters

Silhouette Score: 0.29.

This silhouette score represents the moderate performance of the cluster. It demonstrates that there is a significant overlap between the clusters. It also means that there is a high room for improvement or this algorithm is not a very good approach to classify the text data into different clusters.

## 4.3 Hierarchical Clustering

Hierarchical Clustering is another unsupervised machine-learning algorithm that is used in this study for the classification of text data into different classes. Similar to k-mean, the metric that is used to evaluate the model is the silhouette score. Following are the results of this algorithm on text data

Silhouette Score: 0.38

This result illustrates that the value now is slightly better than the k-mean but still not very good as there is overlap among the clusters. As closer to +1 means well-separated clusters while smaller values means overlap among the cluster.

## 4.4 DBSCAN Clustering

DBSCAN clustering is another unsupervised clustering algorithm that is used in this study to classify tweets into different classes. The aim is same as with k-mean and hierarchical clusters algorithm. Following are the result achieved with this algorithm.

Silhouette Score: 0.45.

The issue with this approach is it found about 53 clusters.

Slightly better result as compared to the k-mean and hierarchical cluster but still not very good enough for clustering. Furthermore, the major concern with this cluster is unlike another algorithm that it is not possible to define the number of clusters in advance. This creates an issue as it classifies the data into 53 different classes. This is the reason its silhouette score is quite better than others as there is not much overlap due to a greater number of clusters.

## 4.5 Feedforward Neural Network

In this part, the result obtained after training of feedforward neural network on two different datasets is mentioned and discussed. Four different metrics have been used to evaluate the model. The FNN model and its architecture are explained clearly in the methodology section. Here the aim is to focus on metrics and the performance of the model. This model is trained on two different datasets.

Apple sentiment (4k tweets)

Airline sentiment (14k tweets)

After necessary data preparation steps, the model has been trained and the results have been received for four different model evaluation metrics. The FNN model was trained on both dataset for a 10 number of epochs. The results achieved on both datasets are shown in table 4.2.

**Table 4.2** *Feedforward Neural Network evaluation metrics*

Metrics	FNN Model (Apple data)	FNN model (Airline data)
Test Accuracy	0.71	0.76
F1-Score	0.69	0.76
Precision	0.69	0.77
Recall	0.71	0.76

The table 4.2 shows the metrics of the FNN model for both datasets. The model shows 71% accuracy for Apple sentiment and 76% for airline sentiment. This means that the model correctly classified these percentages of tweets in the test set. It indicates reasonably a good performance identifying the sentiment. Furthermore, all other metrics including F-1 score, Precision, and Recall in each dataset are quite close to their accuracy. All four metrics for Airline sentiments are higher than the Apple sentiment with the same FNN model. It could be because of the big dataset with more tweets.

Overall the performance of the FNN model is better with airline sentiment data than with Apple sentiment. The differences in performance metrics suggest that

the context, size, and nature of the dataset can significantly affect the model performance

With the augmentation technique, the accuracy was reduced further. This technique is tested for Apple sentiment data only.

Test Accuracy: 0.64

Test F1 Score: 0.59

#### 4.5.1 FNN (Apple Sentiment)

		Negative	Neutral	Positive
True Labels	Negative	150	84	4
	Neutral	48	368	13
	Positive	17	55	22
		Predicted Labels		

*Figure 4.1 Feedforward Neural Network Confusion Matrix (Apple Data)*

##### Negative Class:

The figure 4.1 depicts that 150 tweets were negative and are classified as the same. 84 are incorrectly classified as neutral. 4 incorrectly classified as positive.

##### Neutral Class:

368 tweets were neutral and are classified as the same. 48 tweets were not negative but classified as negative 13 tweets were not positive but classified as positive

##### Positive Class

22 tweets were correctly classified as positive. 17 tweets were not negative but classified as negative. 55 tweets were not neutral but classified as Neutral (False Positives for Negative class)

## 4.5.2 FNN (Airline Sentiment)

**Confusion Matrix**

True Labels	Negative	1583	206	100
	Neutral	171	318	91
	Positive	74	59	326
		Negative	Neutral	Positive
		Predicted Labels		

*Figure 4.2 Feedforward Neural Network Confusion Matrix (Airline data)*

### Negative Class:

Figure 4.2 depicts that 1583 tweets were negative and are classified as the same. 206 were not neutral but classified as neutral. 100 were not positive but classified as positive.

### Neutral Class:

318 tweets were correctly classified as Neutral. 171 tweets were not negative but classified as negative 91 tweets were not positive but classified as positive

### Positive Class

326 tweets were correctly classified as positive. 74 tweets were not negative but classified as negative. 59 tweets were not neutral but classified as Neutral (False Positives for Negative class)

The confusion matrix illustrates that the FNN model performs adequately on both datasets. However, it has few misclassifications and a higher accuracy on the airline sentiment dataset.

## 4.6 Recurrent Neural Network

In this part, the result obtained after Recurrent Neural Network (RNN) training on two different datasets is mentioned and discussed. Similar to FNN, four different

metrics have been used to evaluate this model as well. The RNN model and its architecture are explained clearly in the methodology section. Here the aim is to focus on metrics and the performance of the model. The same two datasets are used here to train the RNN model.

After the necessary data preparation steps, the model has been trained and the results have been received for four different model evaluation metrics.

The RNN model was trained on both datasets for a 4 number of epochs. The results achieved on both datasets are shown in the table 4.3.

**Table 4.3** *Recurrent Neural Network evaluation metrics*

Metrics	RNN Model (Apple data)	RNN model (Airline data)
Test Accuracy	0.56	0.79
F1-Score	0.41	0.79
Precision	0.31	0.79
Recall	0.56	0.79

Table 4.3 shows the metrics of the RNN model for both datasets. The model shows 56% accuracy for Apple sentiment and 79% for airline sentiment. This means that the model correctly classified these percentages of tweets in the test set. It indicates not a good performance identifying the sentiment for Apple sentiment while reasonably good for airline sentiment. Furthermore, all other metrics including F-1 score, Precision, and Recall in each dataset are equal to the accuracy in the case of airline sentiment but low in case of apple sentiment. All four metrics for Airline sentiments are higher than the Apple sentiment. It could be because of the big dataset with more tweets.

Overall, like FNN, the performance of the RNN model is better with airline sentiment data than with Apple sentiment. The differences in performance metrics suggest that the context, size, and nature of the dataset can significantly affect the model performance.

With Augmentation Technique. Testing for one apple sentiment data only. It seems like it doesn't make much difference in terms of accuracy.

Test Accuracy: 0.56

Test F1 Score: 0.40

Figure 4.3 and 4.4 are the confusion matrix for both datasets.

## RNN (Apple Sentiment)

**Confusion Matrix**

True Labels	Negative	0	238	0
	Neutral	0	429	0
	Positive	0	94	0
		Negative	Neutral	Positive
		PredictedLabels		

*Figure 4.3 Recurrent Neural Network Confusion Matrix (Apple data)*

**Negative Class** The confusion matrix of figure 4.3 depicts no tweets are accurately classify as negative. 328 were not neutral but classified as neutral. Nothing is classified as positive.

**Neutral Class** 429 tweets were neutral and classified as the same. Nothing is classified as negative or positive.

**Positive Class** Nothing classified as positive and negative. 94 tweets were incorrectly classified as Neutral (False Positives for Negative class)

## RNN (Airline Sentiment)

**Negative Class** Figure 4.4 depicts that 1647 tweets were negative and are classified as the same 125 were neutral and are incorrectly classified as the same. 117 incorrectly classified as positive.

**Neutral Class** 317 tweets were neutral and classified as the same. 188 tweets were not negative but classified as negative. 75 tweets were incorrectly classified as positive.

**Confusion Matrix**

True Labels	Negative	1647	125	117
	Neutral	188	317	75
	Positive	48	56	355
		Negative	Neutral	Positive
		Predicted Labels		

*Figure 4.4 Recurrent Neural Network Confusion Matrix (Airline data)*

**Positive Class** 355 tweets were positive and classified as the same. 48 tweets were not negative but classified as negative. 56 tweets were incorrectly classified as Neutral (False Positives for Negative class)

## 4.7 BERT

In this part, the result obtained after BERT model training on two different datasets is mentioned and discussed. Similar to above two models, four different metrics have been used to evaluate this model as well. The BERT model and its architecture are explained clearly in the methodology section. Here the aim is to focus on metrics and the performance of the model. The same two datasets are used here to train and evaluate the BERT model.

After the necessary data preparation steps, the model has been trained and the results have been received for four different model evaluation metrics.

Three different variants of the BERT model were trained on both datasets for a 6 number of epochs. The results achieved on both datasets with each variant of BERT are shown in the tables 4.4.

**Bert-base-uncased**

**Bert-large-uncased**

**Roberta**

*Table 4.4 Bert-base uncased evaluation metrics*

<b>Metrics</b>	<b>BERT (Apple data)</b>	<b>BERT model (Airline data)</b>
Test Accuracy	0.8	0.95
F1-Score	0.79	0.95
Precision	0.8	0.951
Recall	0.8	0.951

*Table 4.5 Bert-large-uncased evaluation metrics*

<b>Metrics</b>	<b>BERT (Apple data)</b>	<b>BERT (Airline data)</b>
Test Accuracy	0.56	0.93
F1-Score	0.789	0.93
Precision	0.789	0.93
Recall	0.789	0.93

Table 4.4, 4.5, and 4.6, shows the metrics of the BERT models for both datasets and for all three variants. The bert-based-uncased shows 80% accuracy for Apple sentiment and 95% for airline sentiment. Similarly, bert-large-uncased gave 56% and 93% while roberta-based mode gave 81% with apple sentiment and 94% with Airline sentiment. This means that the model correctly classified these percentages of tweets in the test set. It indicates a good performance in identifying the sentiment for Apple sentiment while more better for airline sentiment in all three variant. Furthermore, all other metrics including F-1 score, Precision, and Recall in each dataset are nearly equal to the accuracy. All four metrics for Airline sentiments are higher than the Apple sentiment. It could be because of the big dataset with more tweets.

Overall, like the other two models, the performance of the BERT model is better with airline sentiment data than with Apple sentiment. Among all three variants, Roberta has higher accuracy for apple sentiment, while bert.based-uncased shows higher accuracy for airline sentiment.

*Table 4.6 Roberta evaluation metrics*

Metrics	BERT (Apple data)	BERT (Airline data)
Test Accuracy	0.81	0.94
F1-Score	0.81	0.94
Precision	0.805	0.94
Recall	0.806	0.94

**BERT (apple Sentiment)**

Confusion Matrix

True Labels	Negative	177	54	7
	Neutral	31	378	20
	Positive	8	39	47
		Negative	Neutral	Positive
		PredictedLabels		

*Figure 4.5 BERT Confusion Matrix (Apple data)*

**Negative Class** The 4.5 depicts that 177 tweets were negative and are classified as the same. 54 were not neutral but classified as neutral. 7 were classified as positive but are not positive.

**Neutral Class** 378 tweets were correctly classified as Neutral. 31 tweets were not negative but classified as negative 20 tweets were not positive but classified as positive

**Positive Class** 47 tweets were correctly classified as positive. 8 tweets were incorrectly classified as negative. 39 tweets were incorrectly classified as Neutral (False Positives for Negative class)

## BERT (Airline Sentiment)

**Confusion Matrix**

		Predicted Labels		
		Negative	Neutral	Positive
True Labels	Negative	1732	95	68
	Neutral	43	1703	40
	Positive	8	15	1803

*Figure 4.6 BERT Confusion Matrix (Airline data)*

**Negative Class** The 4.6 depicts that 1732 tweets were negative and are classified as the same. 95 were not neutral but classified as neutral. 68 were classified as positive but were not positive.

**Neutral Class** 1703 tweets were neutral and classified as the same. 43 tweets were not negative but classified as negative 40 tweets were not positive but classified as positive

**Positive Class** 1803 tweets were positive and classified as the same. 8 tweets were incorrectly classified as negative. 15 tweets were not neutral but classified as Neutral (False Positives for Negative class)

## 4.8 GPT -4

In this part, the result obtained from the GPT-4 model on both datasets. Similar to above models, four different metrics have been used to evaluate this model as well. The GPT-4 model and its architecture are explained clearly in the methodology section. Here the aim is to focus on metrics and the performance of the model.

The model is pre-trained so there is no need to train it again, so it is used to obtain the result received for four different model evaluation metrics. The results achieved on both datasets with GPT-4 models are shown Table 4.7.

*Table 4.7 GPT-4 evaluation metrics*

<b>Metrics</b>	<b>GPT-4 model (Apple data)</b>	<b>GPT-4 model (Airline data)</b>
Test Accuracy	0.8	0.84
F1-Score	0.804	0.85
Precision	0.816	0.87
Recall	0.8	0.85

Table 4.7 shows the metrics of the GPT-4 model for both datasets. The model shows 80% accuracy for Apple sentiment and 84% for airline sentiment. This means that the model correctly classified these percentages of tweets in the test set. It indicates a good performance identifying the sentiment for Apple sentiment while better for airline sentiment. Furthermore, all other metrics including F-1 score, Precision, and Recall in the Apple dataset are equal to the accuracy while in the case of airline sentiment, the precision is a bit higher than the accuracy. All four metrics for Airline sentiments are higher than the Apple sentiment. It could be because of the big dataset with more tweets.

Overall, GPT-4 performance is slightly better with airline sentiment data than with Apple sentiment.

## GPT-4 (Apple Sentiment)

**Confusion Matrix**

		Predicted Labels		
		Negative	Neutral	Positive
True Labels	Negative	189	31	1
	Neutral	27	291	50
	Positive	4	16	67

*Figure 4.7 GPT-4 model Confusion Matrix (Apple data)*

**Negative Class** The figure 4.7 depicts that 189 tweets are correctly classified as negative. 31 are incorrectly classified as neutral. 1 incorrectly classified as positive.

**Neutral Class** 291 tweets were neutral and classified as the same. 27 tweets were not negative but classified as negative 50 tweets were incorrectly classified as positive

**Positive Class** 67 tweets were positive and classified as positive. 16 tweets were not negative but classified as negative. 4 tweets were not neutral but classified as Neutral (False Positives for Negative class)

## GPT-4 (Airline Sentiment)

**Confusion Matrix**

True Labels	Negative	1603	266	20
	Neutral	57	478	45
	Positive	18	68	373
		Negative	Neutral	Positive
		Predicted Labels		

*Figure 4.8 GPT-4 model Confusion Matrix (Airline data)*

**Negative Class** The figure 4.8 depicts that 1603 tweets were negative and are classified as the same. 266 are incorrectly classified as neutral. 20 incorrectly classified as positive.

**Neutral Class** 478 tweets were neutral and classified as the same. 57 tweets were not negative but classified as negative 45 tweets were incorrectly classified as positive

**Positive Class** 373 tweets were positive and classified as positive. 16 tweets were not negative but classified as negative. 68 tweets were incorrectly classified as Neutral (False Positives for Negative class)

## 4.9 Models Comparison

This section contains the comparison of all the models except clustering techniques because the silhouette score for those models are very low. It also means that the performance is not that enough to classify tweets into different classes.

### 4.9.1 Intra BERT comparison for apple sentiment

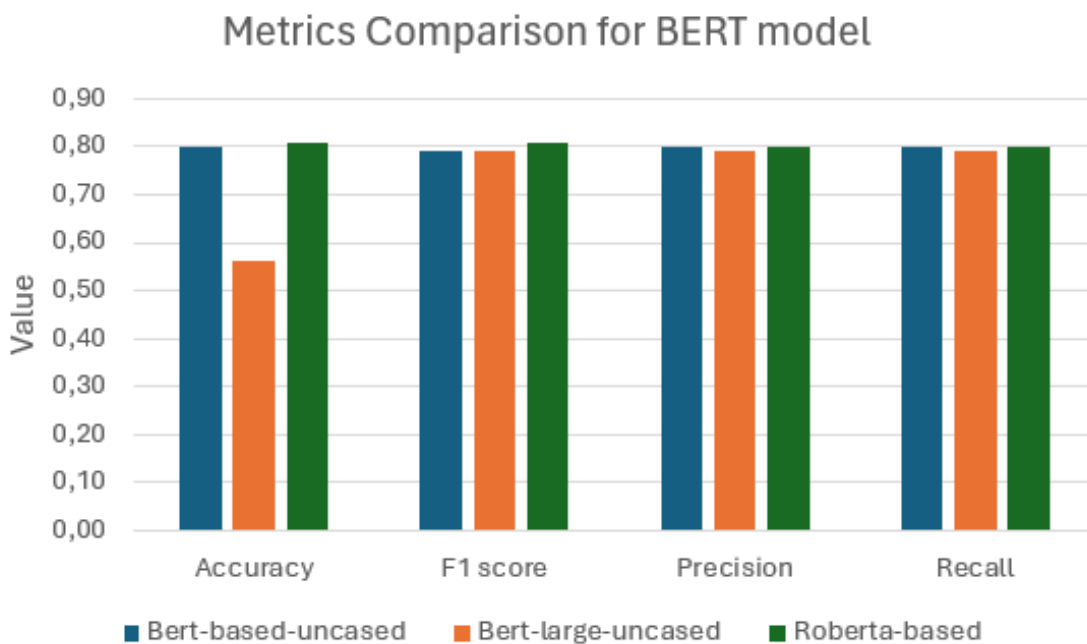
In this intra-bert model comparison, three different variants of BERT are used. These variants are

Bert-base uncased

Bert-large uncased

Roberta-based.

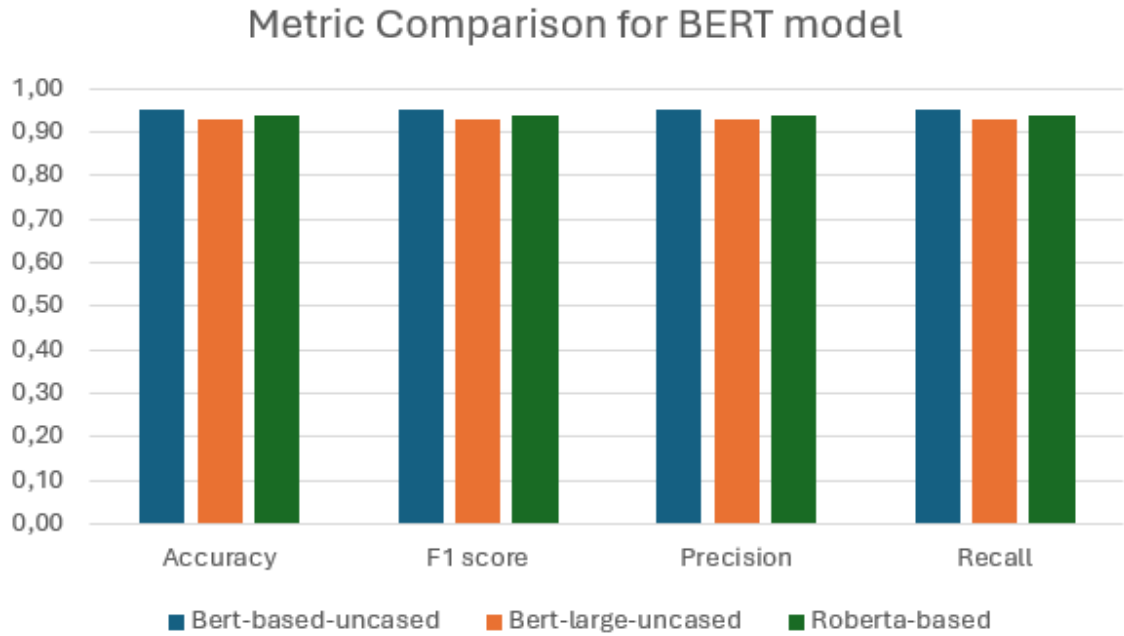
Figure 4.9 is used to compare the metrics of all three variants of the BERT model.



*Figure 4.9 Intra BERT model comparison for apple sentiment)*

Figure 4.9 depicts that comparison of different metrics of all three variants of the BERT model. It is a kind of an intra-model comparison of BERT. It is also clear from the figure 4.9 that Roberta-based is performing better as compared to others competitor in term of accuracy. Roberta-based model is the only model where the accuracy is reached above 80. Furthermore, overall, the green bars which show the value for roberta based have high values for all metrics as compare to others. The worst among the three variants is Bert-large-uncased.

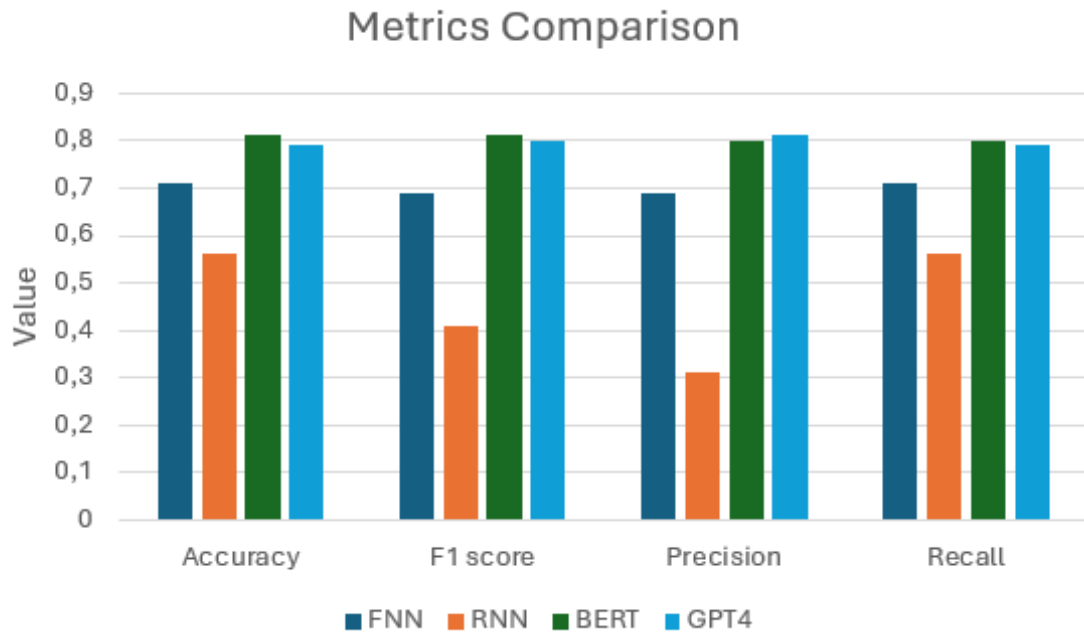
### 4.9.2 Intra BERT comparison for Airline sentiment



*Figure 4.10 Intra BERT model comparison for Airline sentiment)*

Figure 4.10 depicts that comparison of different metrics of all three variants of the BERT model. It is a kind of intra-model comparison of BERT for airline sentiment data. It is also clear from the figure 4.10 that bert-based-uncased is performing better as compared to others competitor in terms of accuracy. Furthermore, overall, the dark blue bars which show the value for bert-based-uncased have high values for all metrics as compared to others. Similar to the apple sentiment case, the worst among the three variants is Bert-large-uncased.

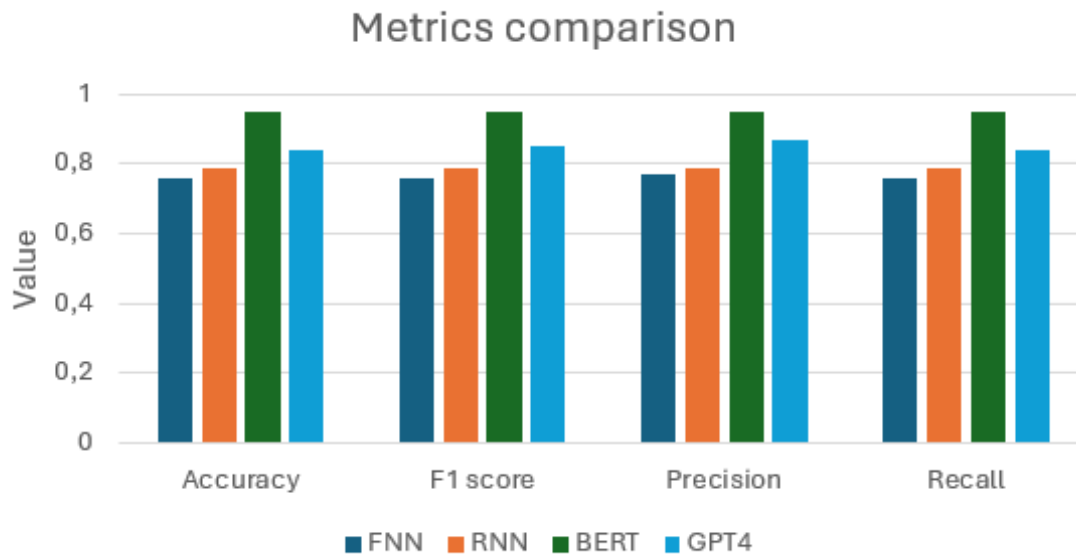
### 4.9.3 All model comparisons for apple sentiment



*Figure 4.11 model comparisons for apple sentiment)*

Figure 4.11 illustrates that comparison of different metrics of all models that have been used in this study. It is a complete summary of the findings of this research study. Figure 4.11 shows that for the classification of text data, BERT is performing better as compared to other competitors in terms of accuracy. The important finding of this research is GPT-4 which is performing quite well. Their metrics value is very close to the BERT. BERT crosses 80 and reaches 81 percent accuracy while GPT-4 has 79 percent accuracy. Though there is a minor difference in values between BERT and GPT-4, but still GPT-4 metrics values are very close to the BERT model.

#### 4.9.4 All Model Comparison for Airline Sentiment



*Figure 4.12 Model Comparison for Airline Sentiment)*

Figure 4.12 shows that the results are similar to what we have for Apple sentiment. The better among all is BERT again. GPT-4 is the second-best model. The Green bar of the BERT model has clear dominance among all other models. The blue bar which represents GPT-4 has second second-best values for all metrics.

### 4.10 Lessons Learnt from the Experiments

In this whole experimental setup, I have used some models from the machine learning domain including K-mean, DBSCAN, Hierarchical clustering, Deep learning domain including FNN and RNN models, BERT, and large language models including GPT.4 for classification of tweets into positive, negative and neutral class. The dataset chosen for this experiment is Apple sentiment and Airline sentiment. The aim is to compare these models' performances in sentiment classification tasks. Multiple key lessons and insights emerged from this experiment. These insights and lessons highlighted the weaknesses and strengths of each method and gave valuable paths to future work related to sentiment data classification.

#### 4.10.1 BERT and GPT-4 superiority

Both BERT and GPT-4 perform very well on sentiment data classification with BERT on the top having GPT-4 on the second top. The difference between the two

is very close in Apple sentiment data where the size of data is smaller as compared to airline sentiment. This further depicts that by understanding the pattern insights text data, the transformer base model is better. The bidirectional approach of BERT models helps the model to capture the context inside the data from both directions. On the other hand, GPT-4 benefits from its pre-training on a vast majority of datasets and architecture making it a robust choice for sentiment data classification tasks.

### **4.10.2 Tradition Approaches Limitation**

This experiment also helps in understanding the limitations of traditional approaches including K-means, Hierarchical clustering, FNN, and RNN, etc. It also depicts the shift of using the latest and more advanced methods for doing sentiment data classification tasks. Specially machine learning approaches like K-mean, DBSCAN, and Hierarchy are not able to classify tweets properly, so these are left out of the final comparison. These algorithms are more suited for tasks having well-defined clusters with proper boundaries rather than subjective and context-dependent classification like sentiment analysis. On the other hand, deep learning algorithms like FNN and RNN are good enough to classify tweets with some accuracy and these are being considered for comparison with other models.

Overall, it gives the lesson of considering the large language models for sentiment classification tasks to avoid pre-processing and training costs. It is also important to consider BERT for classification tasks because of its superior performance due to its transformer architecture and pre-trained on a large corpus of datasets.

## 5 Conclusion and Future Work

This study is about the comparison and classification of sentiment data via LLM, deep learning, and machine learning models. Primarily focuses on using latest model of Openai and compare how it is performing as compared to other model including BERT which is a specialized model for sentiment analysis. For this study several methods like K-mean, DBSCAN, hierarchical clustering for unsupervised domain, FNN and RNN from deep learning domain while BERT and GPT-4 from large language models are chosen for text data classification. After careful analysis and testing different deep learning and LLM models for text data classification, it is now concluded that BERT roberta performs better with 81% accuracy for the Apple dataset, while bert-based-uncased is performing better with Airline data with 95% accuracy than all other models. The only difference between airline and apple dataset is the size of the dataset. Airline datasets contain about four times more tweets than the apple dataset. It provides more training to BERT which leads to more robust and stable model. It also means that it gives more example to the model to train and understand the pattern inside the dataset. The second-best model with slightly less accuracy is 79% for Apple sentiment while 85 for airline sentiment is GPT 4. Models like GPT-4 often used in zero shot or few shot learning which means it just need few sentences to understand the patters. Size of the dataset doesn't affect it performance like BERT. Size helps it to understand more but not the way it is for BERT. Due to this reason, size increase the performance of BERT more than GPT-4. The aim was to compare different models for classifying sentiments which is now done, and BERT has turned out to be the best with GPT-4 as the second-best model. It can also be concluded that LLM models are better for sentiment analysis than deep learning and machine learning models. Unlike machine learning and deep learning models, LLM doesn't need much cleaning. This kind of study and comparison is quite better in the research area as it shows how different models advances with time and how it can be beneficial for industries in research purpose. This kind of study helps other researchers to grow more with more latest models until the best one is found.

**Future work** It is clear from the result and conclusion sections that GPT 4 accuracy is very close to the BERT model. Also, the GPT 4 model is the latest model from Openai which is trained on larger parameters and performance it is also the latest version of other models of Openai like GPT-2, GPT-3, and GPT-3.5. In this research I have compared the latest model of openai with BERT and its accuracy is quite close to Bert, the future model of openai might surpass Bert in terms of accuracy. In the future, this research can be extended by comparing the future model of openai with the current text classification models like BERT.

The bidirectional approach of BERT is also phenomenal with the highest accuracy. Another possible extension of this thesis is to work on the architecture of BERT to make some possible changes for better accuracy. As the architecture of the BERT model was not the scope of this thesis so in future the architecture can be analyzed for some possible use to see how it can further be improved for sentiment classification tasks.

## 6 References

- 1 Anwar, Zeeshan, et al. "Fuzzy Ensemble of Fined Tuned BERT Models for Domain-Specific Sentiment Analysis of Software Engineering Dataset." *PLOS ONE*, edited by Sanaa Kaddoura, vol. 19, no. 5, May 2024, p. e0300279. DOI.org (Crossref), <https://doi.org/10.1371/journal.pone.0300279>.
- 2 Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... Amodei, D. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*. <https://doi.org/10.48550/arXiv.2005.14165>
- 3 Chen, X., Liu, Y. (2019). Deep learning for text data analysis. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(3), e1306.
- 4 Devlin, Jacob, et al. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805*, arXiv, 24 May 2019. [arXiv.org, http://arxiv.org/abs/1810.04805](http://arxiv.org/abs/1810.04805).
- 5 Devlin, J., Chang, M. W., Lee, K., Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171-4186. <https://doi.org/10.48550/arXiv.1810.04805>
- 6 Duan, Ruixue, et al. "Sentiment Classification Algorithm Based on the Cascade of BERT Model and Adaptive Sentiment Dictionary." *Wireless Communications and Mobile Computing*, edited by Jinbo Xiong, vol. 2021, Aug. 2021, pp. 1–8. DOI.org (Crossref), <https://doi.org/10.1155/2021/8785413>.
- 7 Fei, Rong, et al. "Deep Learning Structure for Cross-Domain Sentiment Classification Based on Improved Cross Entropy and Weight." *Scientific Programming*, vol. 2020, June 2020, pp. 1–20. DOI.org (Crossref), <https://doi.org/10.1155/2020/3810261>
- 8 Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., Steinhardt, J. (2021). Measuring Massive Multitask Language Understanding. *ICML*. <https://doi.org/10.48550/arXiv.2009.03300>
- 9 Li, Wenkuan, et al. "An Improved Approach for Text Sentiment Classification Based on a Deep Neural Network via a Sentiment Attention Mechanism." *Future Internet*, vol. 11, no. 4, Apr. 2019, p. 96. DOI.org (Crossref). <https://doi.org/10.3390/fi11040096>

- 10 Liu, Yinhan, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692, arXiv, 26 July 2019. arXiv.org, <http://arxiv.org/abs/1907.11692>
- 11 Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692. <https://doi.org/10.48550/arXiv.1907.11692>
- 12 Michael, Eshun, et al. "Sentiment Analysis and Classification of Ghanaian Football Tweets from the 2022 FIFA World Cup." Indonesian Journal of Electrical Engineering and Computer Science, vol. 34, no. 1, Apr. 2024, p. 497. DOI.org (Crossref), <https://doi.org/10.11591/ijeecs.v34.i1.pp497-507>.
- 13 Munikar, Manish, et al. "Fine-Grained Sentiment Classification Using BERT." 2019 Artificial Intelligence for Transforming Business and Society (AITB), IEEE, 2019, pp. 1–5. DOI.org (Crossref), <https://doi.org/10.1109/AITB48515.2019.8947435>
- 14 Obinwanne, T., Brandtner, P. (2023, August). Enhancing Sentiment Analysis with GPT—A Comparison of Large Language Models and Traditional Machine Learning Techniques. In International conference on WorldS4 (pp. 187-197). Singapore: Springer Nature Singapore. [https://link.springer.com/chapter/10.1007/978-981-99-7569-3\\_17](https://link.springer.com/chapter/10.1007/978-981-99-7569-3_17)
- 15 Pandey, Chaitanya. "redBERT: A Topic Discovery and Deep Sentiment Classification Model on COVID-19 Online Discussions Using BERT NLP Model." International Journal of Open Source Software and Processes, vol. 12, no. 3, July 2021, pp. 32–47. DOI.org (Crossref), <https://doi.org/10.4018/IJOSSP.2021070103>.
- 16 Pennington, Jeffrey, et al. "Glove: Global Vectors for Word Representation." Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, 2014, pp. 1532–43. DOI.org (Crossref), <https://doi.org/10.3115/v1/D14-1162>.
- 17 Plaza, Dwaine, and Lauren Plaza. "Facebook and WhatsApp as Elements in Transnational Care Chains for the Trinidadian Diaspora." Genealogy, vol. 3, no. 2, 2019, p. 15. <https://doi.org/10.3390/genealogy3020015>
- 18 Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9. [https://d4mucfpksywv.cloudfront.net/better-language-models/language\\_models\\_are\\_unsupervised\\_multitask\\_learners.pdf](https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf)

- 19 Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., ... Irving, G. (2022). Scaling Language Models: Methods, Analysis Insights from Training Gopher. arXiv preprint arXiv:2112.11446. <https://doi.org/10.48550/arXiv.2112.11446>
- 20 Smith, M. R., Martinez, T. (2016). A comparative evaluation of curriculum learning with filtering and boosting in supervised classification problems. *Computational Intelligence*, 32(2), 167-195. <https://doi.org/10.1111/coin.12047>
- 21 Wang, Chuantao, et al. "Deep Learning Sentiment Classification Based on Weak Tagging Information." *IEEE Access*, vol. 9, 2021, pp. 66509–18. DOI.org (Crossref), <https://doi.org/10.1109/ACCESS.2021.3077059> .
- 22 Wang, Tianyi, et al. "COVID-19 Sensing: Negative Sentiment Analysis on Social Media in China via BERT Model." *IEEE Access*, vol. 8, 2020, pp. 138162–69. DOI.org (Crossref), <https://doi.org/10.1109/ACCESS.2020.3012595>.
- 23 Williams, C., Davis, E. (2018). Unsupervised text data labeling using clustering techniques. *Proceedings of the 15th International Conference on Natural Language Processing*, 123-134.
- 24 Wu, Yichao, et al. Research on the Application of Deep Learning-Based BERT Model in Sentiment Analysis. arXiv:2403.08217, arXiv, 12 Mar. 2024. arXiv.org, <http://arxiv.org/abs/2403.08217>.
- 25 Xiang, Minhao, et al. "A Study on Online Health Community Users' Information Demands Based on the BERT-LDA Model." *Healthcare*, vol. 11, no. 15, July 2023, p. 2142. DOI.org (Crossref), <https://doi.org/10.3390/healthcare11152142>.
- 26 Xiong, W., Wu, L., Alleva, F., Droppo, J., Huang, X., Stolcke, A. (2018). The Microsoft 2017 Conversational Speech Recognition System. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5934-5938). IEEE. DOI.org <https://doi.org/10.1109/ICASSP.2018.8461870>
- 27 Skenderi, E., et al. (2024) INCEPT: A Framework for Duplicate Posts Classification with Combined Text Representations, *ACM Transaction on the Web* <https://doi.org/10.1145/3677322>.