ERASMUS UNIVERSITY ROTTERDAM

Erasmus School Of Economics

Master Thesis in Data Science and Marketing Analytics

"Assessing the Impact of Transitioning to a Freemium Subscription-based Model on Public Sentiment and The Use of Large Language Models in Sentiment Analysis"

Student Name: Alexandros Stavropoulos

Student Number: 643414

Supervisor: Jeffrey Durieux

Second Assessor: Radek Karpienko

Date of final version: July 2023

Abstract

This study investigated the potential impact of implementing a freemium business model with a monthly cost for advanced features, on the public opinion of Artificial Intelligence (AI) platforms. The study primarily focused on the ChatGPT platform, developed by OpenAI. Additionally, the study also assessed the effectiveness of Large Language Models in the task of sentiment analysis compared to other traditional machine and deep learning methods. Furthermore,the study explored the potential benefits of combining an LLM with a Convolutional Neural Network. Moreover, the BERTopic model was employed in order to identify the factors that influenced the public opinion towards ChatGPT. The outcome of this study demonstrated that the transition to the freemium model did not result in a permanent effect on public sentiment, but rather caused only to temporal fluctuations. This suggests that the long term perception of AI platforms by the public can not be affected with a business model transition, showcasing a general adaptability and flexibility among the users in changes like this. For the task of sentiment analysis, various models were examined including models like Naive Bayes classifier, Support Vector Machine, BERT, RoBERTa and DistilBERT. The RoBERTa model outperformed all others, achieving an accuracy of approximately 85%. The combination of DistilBERT model with the Convolutional Neural Network model did not enhance the performance of the DistilBERT model. The public sentiment was influenced by a wide range of topics. Positive sentiments were primarily associated with topics like the use of ChatGPT in everyday practical matters and the utility of it in special occasions, while the negative ones were related to regulatory implications and ethical considerations of the platform. The findings of the study offer significant insights from both managerial and academic standpoints, thereby contributing to the existing body of literature on freemium business models and sentiment analysis.

Contents

List of Tables

List of Figures

1 Introduction

Within the dynamic environment of digital services, the Freemium business model has emerged as a strategy to provide online services to users. Freemium refers to the combination of a free subscription to the platform that gives the user some basic but limited features, and a premium subscription with a monthly fee that provides the user all the advanced features of the platform without any limitation. This model is widely utilized as business model across various digital platforms, including music streaming services and numerous Artificial Intelligence (AI) platforms.

With the growing development of AI of recent years and the widespread adoption of AI-based platforms an increasing demand for subscription-based AI platforms and AI services that provide advanced features and capabilities is recorded (1). However, while these platforms offer many benefits, there are also concerns about their impact on society, such as issues related to data privacy, bias, and job displacement (2). Therefore, it is important to understand people's opinions and sentiments towards these platforms, in order to identify areas for improvement and to address potential challenges. However, it is quite unclear whether consumers are willing adopt AI services on their day to day tasks and also their willingness to pay for these platforms and how the launch of a premium subscription would affect their attitudes.

Social media and especially Twitter has recently become the main tool for individuals to share and express their thoughts on a wide range of topics ranging from news, social events and matters, celebrities, products and trending services (3). Companies can gain a huge advantage in decision making from understanding their costumers or potential customers opinion about their products or their services (4). This study aims to perform sentiment analysis on subscription-based AI platforms based on people's opinion on social media. The main goal is to identify what affect can have an implementation of a freemium business model, what topics affected the aforementioned sentiment and a comparison of the sentiment analysis methods that companies can use.

In order to perform the sentiment analysis the study will focus on the recently launched AI platform ChatGPT. ChatGPT is a chatbot platform that uses AI in order to understand and interact with the end user. It was launched in November 2022. In February 2023, OpenAI, the provider of the platform, decided to make a transition from a completely free online platform into using a Freemium business model. In this context, some crucial question arise:

1. *How can public opinion of AI platforms be affected by the introduction of a freemium*

business model with a monthly cost for advanced features ?

2. *What are the topics that affect how people feel about OpenAI and ChatGPT ?*

3. *Is there a difference in classification accuracy for sentiment analysis between Machine - Deep Learning models and Large Language Models(LLM) ?*

4. *Could a combination of a LLM with a Deep learning model provide superior classification accuracy for sentiment analysis ?*

The insights from this study could potentially be useful not only to AI platform providers but to all tech companies that are willing to launch a freemium or even a premium subscription for their online platform and make a transition from a free platform.

In the following sections of this study, a detailed examination of the freemium business model, sentiment analysis, and topic modeling is initially conducted through a review of the relevant academic literature. Then, the selected methodologies will be explained in depth. Afterwards, a dataset description will be presented, followed by an initial exploratory analysis. The following sections, investigate the freemium model's effect on public sentiment over time, with a comparative analysis of different sentiment analysis models and topic modeling.The final sections of the study consist of a discussion of key findings, implications and study limitations.

2 Literature Review

The literature review of the study is centered on three distinct sections, with the aim of gaining a more profound comprehension of the research inquiries presented. The three sections consist of the adoption of a freemium business model, the utilization of machine learning algorithms, deep learning models, and large language models in sentiment analysis, and lastly, the application of topic modeling to identify factors and subjects that impact sentiment and feedback.

2.1 Freemium Business Model

Freemium business model has gained a significant traction over the course of recent years, across a wide range of digital industries and platforms, most importantly on mobile gaming applications and music services. The academic literature associated with the utilization of freemium model is not quite extensive and diverse as the day of conducting this study. During their study Hamari et al. (5), focused on examining the correlation between the value that consumers and users receive and their tendency to use freemium services. They suggested that the higher the perceived value of the freemium service, user's intention to use the service overall raises with the implementation of freemium services. In the same context, Niemand et al. (6), showed that the freemium services oppose greater value for the users than traditional business models like fully premium paid, or fully free models. The reason for is that the the value of a paid service can significantly impact and influence the willingness of a user to pay for advanced features of a service. On the other hand, in their own study Wagner et al. (7), examined the potential of the free version of the freemium model to act like a promotion for the paid version of the model. The results of the study showed that the free version does not act like a promotion to the paid version and therefore it can not influence the willingness of users to convert to the paid version. Furthermore, Mäntymäki et al. (8), examined the factors that can potentially influence the user's willingness to transition from the free version to the paid version of the freemium model. The findings of the study showed that perceived value of the paid version along witch the price tag can influence users choice. Moreover, the findings suggested that the users decision to maintain their paid subscription could be influenced by the features that the paid version can provide to them. Finally Josimovski et al. (9), analyzed the advantages and disadvantages of implementing a freemium business model and what impact can have an act like this. The findings of the study revealed that a freemium model can draw in a substantial customer base, educate users to new feature and finally test

new application alongside with the paid version.

2.2 Sentiment Analysis

Sentiment analysis is defined as the process of extracting the emotion and the attitude of public on a specific event, product or a person. Sentiment analysis is a primary objective within the field of Natural Language Processing (NLP). Recent papers on sentiment analysis have shown that it is possible to extract not only the positive or negative sentiment of people, but also a broader range of emotions and behaviors based on their reviews, tweets, or other forms of texts. Sentiment analysis can be achieved via different techniques, most notably with machine learning and deep learning models. In 2020, Yadav and Vishwakarma (10), attempted to classify different approaches for sentiment analysis based on the model used for the analysis. The approaches are the following :

1. Lexicon-based approaches, to extract the sentiment at a world level of a sentence by using a sentiment dictionary

2. Machine learning approaches by training machine learning classifiers to classify sentences based on their sentiment

3. Deep learning approaches, by training more complex models such as neural networks to better extract the sentiment of a sentence.

4. Large Language Models, by fine tuning, already pre-trained models on large amount of text.

In 2014 Silva et al. (11), performed experiments in order to compare lexicon based approaches with machine learning approaches in order to classify various tweet datasets into positive and negative, The results showed that every machine learning method outperformed the lexicon based approaches but the combination of lexicon and machine based methods improved the accuracy even more. The machine learning models that were used during the experiments were Multinomial Naive Bayes, SVM, Random Forest and Logistic Regression. A sentiment analysis of Twitter posts containing movie reviews was conducted by utilizing Naïve Bayes and Support Vector Machine (SVM) models,by Amolik et al. (12) in 2016. Results showed that the SVM model outperformed the Naïve Bayes model significantly. Also the study concludes that by increasing the training data the accuracy increases. On their study Weber and Syed (13), performed sentiment analysis on 71239 tweets with the use of only machine learning

algorithms. During the study the authors used SVM, Logistic regression, Multinomial Naïve Bayes, Bernouli Naïve Bayes, Decision Trees and ADA Boost. In their study, SVM showed the best performance in classifying tweets based on their sentiment.

Tammina and Annareby (14), in their paper demonstrated a Convolutional Neural Network outperforms significantly Random Forest and Naïve Bayes in predicting the sentiment of product reviews from Amazon and this holds also for IMDB movie reviews. In 2017 Ramadhani and Goo (15), executed a comparative research between Deep Neural Network (DNN) and Multi – Layer Perception model on performing sentiment analysis of Tweets in 2 different languages. The outcome of the research was that the DNN model had an accuracy 75.03% in predicting correctly the sentiment of a tweet, while the MLP model had an overall accuracy of 52.6% In 2021 Rui and Man (16), suggested the utilization of a cutting-edge, pre-trained deep learning model known as BERT and also a Convolutional Neural Network (CNN) in order to perform sentiment on public data set that consists of hotel reviews and both models showed really good results compared to traditional techniques such as Word2Vec. In the same sense, in 2020 Guo and Zhang (17), performed a sentiment analysis on commodity reviews using the bidirectional encoder representation from transformers (BERT), CNN and also a combination of the two. The results showed that while both models alone performed great the combined method outperforms both of them by 14.4% and 17.4% respectively. Furthermore, in 2021 Bozanta et al. (18), propose two transformer-based LLM : BERT and the universal language model fine-tuning (ULMFiT). Both models are pre-trained on large corpora of text data and fine-tuned on StockTwits data for sentiment analysis. The performance of their models on a dataset of over 11,000 StockTwits messages, annotated with sentiment labels (positive, negative, or neutral). They report an accuracy of 83.42% for BERT and 79.31% for ULMFiT, outperforming several baseline models such as logistic regression, SVM, and CNN. In 2023, Suhartono et al. (19), proposed a deep learning approach for extracting the sentiment of drug product reviews. The proposed method architecture used a combination of glove Word embeddings and a CNN model. The model was compared with the state of art model RoBerta. The findings showed that while RoBerta outperforms the proposed model in the initial phase of training, the proposed combination of CNN model and Word2Vec resulted in superior performance while testing the models on a separate test set. In the following Table 1, a summary of key sentiment analysis studies is presented. The Table is focused specifically on the models that showed the best classification accuracy.

	Authors	Paper Name	Model Used	Classification Accuracy
2014	Silva et al.	Tweet sentiment analysis with classifier ensembles	ML+Lexicon-based	76.99%
2016	Amolik et al.	Twitter sentiment analysis of movie reviews using machine learning techniques	SVM	75%
2017	Ramadhani et al.	Twitter sentiment analysis using deep learning methods	DNN	75.03%
2019	Weber et al.	Interdisciplinary optimism? sentiment analysis of Twitter data	SVM	83%
2020	Tammina et al.	Sentiment analysis on customer reviews using convolutional neural network	Convolutional Neural Network	74%
2020	Guo et al.	A commodity review sentiment analysis based on BERT-CNN model	BERT+CNN	84.5%
2021	Rui et al.	Sentiment analysis algorithm based on BERT and convolutional neural network	BERT+CNN	90.5%
2021	Bozanta et al.	Sentiment analysis of StockTwits using transformer models	BERT	83.42%
2023	Suhartono et al.	Deep neural networks and weighted word embeddings for sentiment analysis of drug product reviews	CNN+Word2Vec	91%

Table 1: Summary of Key Sentiment Analysis Studies

2.3 Topic Modelling

The field of topic modelling is currently experiencing a rapid development in the field of Natural Language Processing. The main goal of topic modelling is to reveal patterns within a certain set of documents, for example a set of tweets, in order to perform document clustering and information retrieval. (20). In 2023 Taghandiki et al. (21) overview in their paper the most commonly used algorithms and tools in topic modeling such as Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization(NMF). In a comparative study of various

topic modelling algorithms and techniques that were applied on Twitter data, Egger et al. (22), evaluated the performance of four models, namely LDA, NMF, Top2Vec and BERTopic. The results of the research showed that BERTopic and NMF models outperformed all other models at correctly identifying the topics of the tweets. In the same context, Kherwa and Bansal (20), also conducted a thorough comparison of different topic modelling methodologies. The study examined the following models, LDA, PLSA, Dynamic Topic Model, Syntactic Topic Model and Multilingual Topic Model. The researchers emphasized on the significance of the coherence as an indicator of the degree of association among words within a particular topic. Furthermore, Muchene and Safari (23), in their paper proposed a topic modelling framework divided into two phases. During the first phase LDA was used to obtain the topic probabilities of each of the documents. The next phase uses hierarchical clustering using the Hellinger distance in order to group each document based on their topic. In 2022, George et al. (24), proposed an interesting framework for topic modelling by incorporating clustering based on dimensionality reduction alongside with BERT and LDA model. The results of the study showed that the clustering BERT-LDA based approach can increase the effectiveness of topic modelling with more coherent topics in contrast to other topic modelling approaches.

2.4 Literature Contribution

From a business and marketing perspective, this study extents the literature on the freemium business model by gaining insights from people's opinion about transitioning from a completely free online platform to a freemium business model. It further investigates whether this transition affects the overall sentiment of users.

Although sentiment analysis as a topic has already been extensively explored in recent years, most studies have focused mostly on product reviews. Sentiment analysis of tweets towards subscription based platforms as a paid services and especially AI platforms, yet remain to be discovered. Furthermore, most of the past research on the topic has focused on machine learning methods, with fewer studies examining deep learning models and LLM. This study's focuses on comparing machine learning methods, deep learning methods and LLM and a combination of a LLM and a deep learning method. The ultimate goal of the comparison is to determine if there is a significant difference between them and how the accuracy of sentiment extractions can be improved.

3 Methodology Overview

In this section, a detailed explanation of the methodologies used in this study will be carried out. The analysis begins with an examination of the Ordinary Least Squares (OLS) regression in Interrupted Time Series analysis in section 3.1. Section 3.2 explores machine learning algorithms in sentiment analysis, specifically Support Vector Machine and Naive Bayes. Moreover, Section 3.3 examines the use of Convolutional Neural Network in sentiment analysis. Finally, Section 3.4 provides an examination of Large Language Models like BERT. RoBERTa, DistilBERT and BERTopic models.

3.1 Ordinary Least Squares Regression & Interrupted Time Series

The initial model that was employed in this study is OLS regression method. The OLS regression is a statistical technique that, it is widely used to estimate and determine the associations between variables (25). In order to estimate the association between the variables the model uses the minimization of the sum of squared differences between the observed and predicted values of the dependent variable, which are modeled as a linear relationship (26).

The OLS regression model can be expressed in a general form as follows:

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon$

Where:

- Y is the dependent variable,
- β_0 stands for the intercept,
- β_1 to β_n represent the coefficients with the independent variables X_1 to X_n,
- ε denotes the error term.

In the context of assessing the impact of a disruption or an intervention in a time series data, the OLS regression model can be modified in order to estimate the effects of a specific event before and after it in the time series (27), (28).

The OLS model for the analysis of an interrupted time series is as follows:

$Y_t = \beta_0 + \beta_1 t + \beta_2 X_t + \beta_3 X_t t + \varepsilon_t$

Where:

- Y_t stands for the dependent variable measured at a specific time point t,
- t stands for the elapsed time,

- X_t is an indicator for pre-intervention ($X_t = 0$) or post-intervention ($X_t = 1$),
- β_0 estimates the baseline level of the outcome,
- β_1 estimates the baseline trend of the outcome,
- β_2 estimates the change in level immediately following the intervention,
- β_3 estimates the change in trend following the intervention,
- ε_t is the error term at time t.

The primary parameters of concern of the above model are β_2 and β_3, which can indicate the direct effect of the intervention and the alteration in the trend following the event of examination of the analysis, correspondingly (29).

The assumptions underlying the OLS regression, namely linearity, independence, homoscedasticity, and normality of the data, are also applicable in the context of the interrupted time series analysis. However, as suggested by Box et al. (30), if the assumptions of the OLS model are not true, it is important to use alternative methodologies or transform the data so that the assumptions are met. For instance, if the residuals do not follow a normal distribution, the dependent variable can be transformed by using a logarithmic transformation. In the same context, If the residuals exhibit auto correlation, it is suggested to employ and use other techniques such as auto-regressive integrated moving average (ARIMA) models.

The rationale behind implementing the ITSA OLS model in this study is to employ a statistical approach for accessing the significance of the changes on the public sentiment after the implementation of the freemium business model on the ChatGPT platform. Additionally, this model is able to consider time dependent trends in sentiment that were already changing prior to the introduction of ChatGPT Plus. Furthermore, the model is able to examine the interaction effects resulting from the combined introduction of ChatGPT PLus and GPT4.

3.2 Machine Learning Algorithms

3.2.1 Support Vector Machine

The Support Vector Machine (SVM) is a widely utilized supervised learning algorithms as a probabilistic binary classifier(31).

An SVM algorithm tries to classify data by utilizing decision planes in order to establish boundaries within the data points. A decision plane or a hyperplane separates data objects based on their respective class. The hyperplane is chosen in such a way that the distance

between the hyperplane itself and data points from both classes is maximized (31). The closest points to the hyperplane are called the support vectors, and they define the area that surrounds the hyperplane in order to define the margin that separates the two distinct data classes. The ultimate goal of an SVM algorithm is to maximize the margin of the between classes and therefore to create a highly robust decision boundary (31). The decision boundary, can be mathematically represented as follows:

$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$

In the above expression:

- $\mathbf{w}^T$ is the weight vector, witch determines the direction that the decision boundary will follow.
- $\mathbf{x}$ stands for the input vector withe all the features of a given data point.
- b is the bias

The above function $f(\mathbf{x})$ provides with the output of the SVM algorithm for a given data point. If the outcome of the algorithm is positive then the data point is classified to the baseline classification class, while when the outcome is negative the data point is classified to the alternative classification class (31). It is obvious that the SVM algorithm is primarily built for binary classification. When there are more than two classification class, the model must be used with a different approach.

In order to handle multi class classification, the SVM model can be combined with the one-vs-rest(OvR) technique. The OvR strategy implies employing one SVM model for each class individually (32). Based on this approach, the data points of a particular classification class are considered as the positive samples, while the data points of all the other data points are considered as negative samples. When the model tries to classify a data point, it uses all k classifiers. The classifier that it is responsible for assigning the data point to a specific class, is the one that yields the highest output score (32).

The efficiency of SVM model is dependent on the selection and optimization of the hyper-parameters of the model. The hyperparameters of the model play a crucial role in determining the model's complexity. The optimization of the hyperparameters aims to find a balance between bias and variance (31). The two main hyperparameters of SVM model are the cost parameter C and the kernel parameter .

The cost parameter C is responsible for deciding the balance between the margin of the decision function and the process of classification the training example (33). An increase in

the value of the parameter C results in a smaller decision margin that could effectively learn the training example but it can also lead in overfitting, in which the model shows strong performance in the training phases, however it fails to properly generalize the data in order to perform on unseen data. On the other hand, a smaller value of the C parameter, can prevent overfitting but could potential lead to a lower classification accuracy (33).

When the SVM model is used with the default Radial Basis Function (RBF) kernel the kernel parameter is responsible for determining how much an individual training examples affects the decision function(31). A small value indicates that the impact of one training example covers a much larger area of the high dimensional space, and as a result the decision boundary is smoother, while in contrast a higher value can lead to a more complex decision boundary due to the fact that the influence of one training example is not as strong (33). Additional types of kernels that can be used in SVMs include linear, polynomial, and sigmoid kernels (34).

The optimization phase of the hyperparameters consists of various techniques, including grid search and cross validation. These techniques asses different combination of the hyper-parameters on a validation set, and afterward selecting the combination that resulted in the highest performance of the model.(33)

It is important to note that, in order to use the SVM model with textual data for sentiment analysis it is vital to transform the data into numerical format. Common methodologies for this transformation are Bag of Words or TF-IDF.(33)

3.2.2 Naive Bayes

Naive Bayes algorithm is considered a probabilistic algorithm that makes use as foundation the principles of Bayes's Theorem. The term 'naive' is used due to the fact that the model makes the assumption that all features of certain group of data points are independent to the occurrence of any other feature, regardless the relationship between them (35).

As stated the model is founded upon the Theorem of Bayes. This indicates that the likelihood of an event occurring can be determined by a preexisting knowledge of the conditions that could potentially be correlated with the event.

In a mathematical context, it can be stated as follows:

$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$

Where:

- $P(A|B)$ is the posterior probability,

- $P(B|A)$ is the likelihood,
- $P(A)$ is the class prior probability, and
- $P(B)$ is the predictor prior probability.

The above formula can be translated within the framework of the Naive Bayes classifier, with A representing the prediction class while B representing another prediction class. The classifier tries to assign a certain data point to a specific class by using the mathematical model in order to calculate the probability of the data point belonging to each class (36).

As McCallum et al. (37), stated in their paper Naive Bayes classifier is widely favored and used with high-dimensional datasets and textual data for tasks like text classification. On the other hand, the assumption of feature independence can be a strong limitation for the model. In many practical scenarios, the aforementioned assumptions can result in not optimal performance. Moreover, the model could perform poorly with datasets where the probability of dealing with a specific combination of features may be extremely rare (38). The most important limitation of the model is the zero frequency problem. When a specific class and a specific feature do not co exist in the training data set, then the probability will be estimated as zero. This requires added techniques to overcome this issue such as Laplace smoothing.

The Naive Bayes classifier can be used with three different distributions. To begin with the Naive Bayes model uses Multinomial distribution with features vectors representation or Bernoulli distribution with binary term occurrence features for text and document classification. Furthermore, the Gaussian distribution is used by the model when attempting to classify features that follow a normal distribution (39).

3.3 Deep Learning

3.3.1 Convolutional Neural Network

A Convolutional Neural Network (CNN) is a deep learning model that is widely used primary for image recognition and classification and secondly for sound or text classification. CNNs were were inspired from the structural organization of the visual cortex in animals (40). The study that was conducted by researchers revealed that the neurons situated in the visual cortex exhibit a restricted local receptive field, indicating their responsiveness exclusively to visual stimuli confined within a specific region of the visual field. The influence of the visual cortex is visible in the architecture of CNNs models, where the arrangement of the neuron

layers is organised and structured in three dimensions, namely width, height, and depth (40). CNNs are engineered in such a way to understand and acquire spatial hierarchies of features of the input data in an adaptive and automatic manner (41).

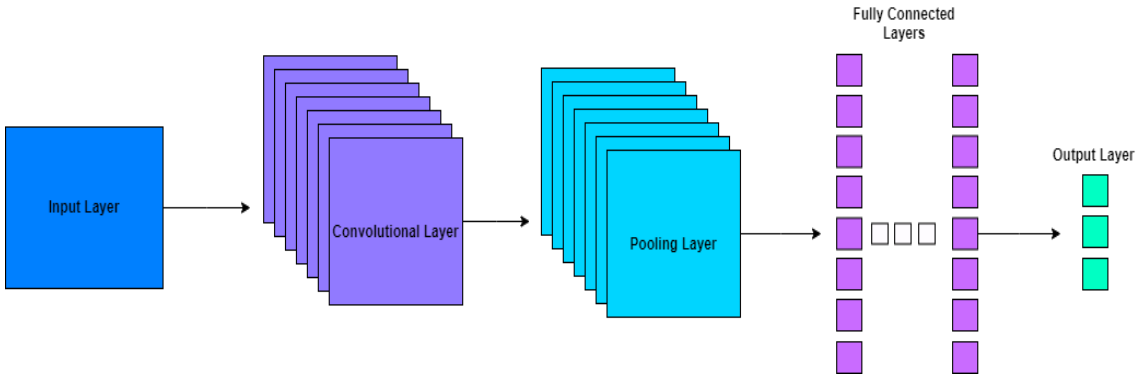

Figure 1: Typical CNN Structure.

While there not a specific way of creating a CNN model, in general a CNN structure consists of multiple layers, namely the input layer of the data, a convolutional layer,a pooling layer,a fully connected layer, and finally an output layer (42). A typical CNN structure can be seen in Figure 1. The input layer of a CNN model is designed to accept 2 dimensional objects such as image or speech or text input.

The convolutional layer is the fundamental component or layer of a CNN model. This layer consist of a wide collection of filters named kernels that are fully learnable (43). The main task of the convolutional layer is to perform the convolution operation during the forward pass. The convolution is done by traversing the input volume in both width and height dimensions, and calculate the dot product between the filter entries and the corresponding input entries. Consequently, the application of the kernel produces a two-dimensional activation map. In this initial layer, the network can develop the capacity to understand and perceive specific features such as visual attributes when the input is an image. As the network proceed to the next layers, its ability to recognize more specific patterns of the input data increases (43). The convolution operation can be defined by the following formula given an input matrix I and a kernel K as:

$(I * K)[i, j] = \sum_m \sum_n I[m, n] \cdot K[i - m, j - n]$

Where:

- $(I * K)[i, j]$ is the element at the i-th row and j-th column of the output matrix,
- $I[m, n]$ is the element at the m-th row and n-th column of the input matrix,

- $K[i-m, j-n]$ is the element at the $(i-m)$-th row and $(j-n)$-th column of the kernel.

Another important layer of a CNN model is a pooling layer. The main purpose of a pooling layer is to decrease the spatial dimensions of the input matrix (44). The pooling operation is extremely useful for the model to extract more prominent characteristics of the data and also helps the training of the model to be more efficient. Mathematically the pooling operation is described as:

$Y[i, j] = \max(I[i : i + k, j : j + k])$

Where:

- $Y[i, j]$ is the element at the i-th row and j-th column of the output matrix,
- $I[i : i + k, j : j + k]$ represents the elements within the pooling window in the input matrix,
- max is the operation that selects the maximum value within the pooling window,
- k is the size of the pooling filter.

For sentiment analysis CNN models requires the pre processing of the textual input data, and transforming it into a numerical format. This transformation is most frequently done by utilizing techniques like word embeddings such as Word2Vec or GLove (45). After that the model is trained on the aforementioned numerical data. The output layer of a CNN model that is designed for sentiment analysis with multi class classification, makes use of a Softmax function. The Softmax function is being used in order to convert the output into a probability distribution for the targeted classes. The softmax function is defined as:

$S(y_i) = \frac{e^{y_i}}{\sum_j e^{y_j}}$

Where:

- $S(y_i)$ is the softmax function applied to the i-th output unit,
- y_i is the i-th output unit of the network,
- e^{y_i} is the exponential of the i-th output unit,
- $\sum_j e^{y_j}$ is the sum of the exponentials of all output units,
- and i and j are indices of the output units.

3.4 Large Language Models

The following section aims to provide an overview of the four LLMs that were used in this study, namely BERT, RoBERTa, DistilBERT. The focus is to define their distinct characteristics, architectures and parameters. Furthermore, in this section BERTopic will be examined, which utilizes LLM for the purpose of topic modeling. Table 2 provides a brief summary of the main characteristics of the models.

	Variant	Trans. Blocks	Hidden Size	Attn. Heads	Params	Next Sentence Pred.	Distillation
BERT	Base	12	768	12	110M	Yes	No
BERT	Large	24	1024	16	340M	Yes	No
RoBERTa	Base	12	768	12	125M	No	No
RoBERTa	Large	24	1024	16	355M	No	No
DistilBERT	Base	6	768	12	66M	Yes	Yes

Table 2: Extended Overview of Large Language Models

3.4.1 Bidirectional Encoder Representations from Transformers (BERT)

The Bidirectional Encoder Representations from Transformers (BERT) model was introduced by the researchers at Google AI Language in 2018 (46). Since its release, BERT has demonstrated exceptional capabilities in the field of natural language processing tasks including sentiment analysis.

The BERT model is created based on the Transformer model, that was first presented in 2017 by Vaswani et al (47). It relies its performance entirely on drawing global dependencies between the input and output data exclusively with the use of an attention mechanism. This attention mechanism also known as "Scaled Dot-Product Attention" mechanism tries to evaluate the relevance of each word with the surrounding context of the word that is being evaluated. The process of the attention mechanism involves calculating the dot products between the query of input data and all keys and then it is divided by the square root of the dimensions of the queries and the keys (47). Both the encoder and the decoder of the model, are constructed by using stacked self attention and point wise linked layers.

In contrast to other related models, BERT has been engineered in such a way that it takes into consideration the whole contextual information of word. A crucial role to this, plays the employment of a method known as Masked Language Model(MLM). The MLM method randomly masks out a certain number of the input tokens in order to predict the original word excursively based on the surrounding context and learn from the mistakes (46). This strategy allows the model to better understand the surrounding context and somatic

interpretation of the words with much greater precision (48). Another key objective of BERT model alongside the MLM is the Next Sentence Prediction (NSP). Durin the training phase the BERT model is provided as input pairs of sentences and tries to identify if the second sentence is the subsequent sentence in the original document (46). In general NSP aids BERT to identify correlation between two sentences, which is extremely useful for NLP tasks like text generation and question answering.

The original BERT model has two distinct variation, the BERT Large and BERT Base and both of them are pre trained on large collection of textual data. The dataset used for training BERT model consisted of more than 3.3 Billion words. While the two variations of BERT model share the same idea behind their architecture that can be seen in Figure 2, they differ in some key parameters.

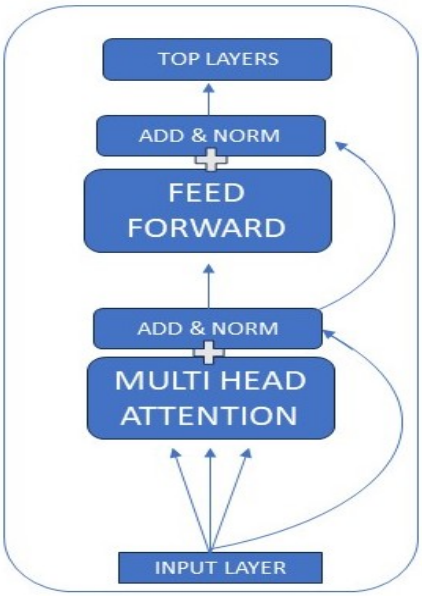

Figure 2: BERT Model General Architecture

The BERTbase model consists of 12 transformer blocks, 768 hidden size layers, 12 attention heads, 110 million parameters and it was trained for 4 days by 4 TPUs as processing unit. On the other hand BERTlarge model consist of of 24 transformer blocks, 1024 hidden size layers, 16 attention heads, 340 million parameters and it was trained for 4 days by 16 TPUs as processing unit. Both versions of the BERT model can be fine tuned by adjusting the top and final layers in order to effectively carry out a specific NLP task (46).

3.4.2 Robustly Optimized BERT Pretraining Approach (RoBERTa)

Robustly Optimized BERT Pretraining Approach (RoBERTa), is considered a variant of BERT model and has been developed by Facebook AI in 2019 by Liu et al. (49). The primary goal of the developers was to find a way to enhance the training process of BERT model and optimize the its hyper parameters. As RoBERTa is a variant of BERT model incorporates the same bidirectional self attention mechanism that Vaswani et al. (47) presented in 2017.

There are some important modifications on the hyperparameters of BERT model. The most important one is the elimination of BERT's Next Sentence Prediction objective, and the training phase is being conducted with significantly larger mini batches and learning rates. RoBERTa has been trained on extended variation of the BookCorpus dataset, which is 10 times larger than the dataset BERT was trained on.

RoBERTa follows exactly the same model architecture of BERT. As already described this architecture comprises of several transformer layers that are equipped with multi head self attention mechanisms and a fully connected feed forward network. The model accepts as its input a series of words and generated a number of vectors as its output.

In the same manner with BERT, RoBERTa has two distinct variation, namely RoBERTa base and RoBERTa large. The RoBERTa base model consists of 12 transformer blocks, 768 hidden size layers, 12 attention heads and 125 million parameters. On the other hand RoBERTa large model consist of of 24 transformer blocks, 1024 hidden size layers, 16 attention heads and 355 million parameters. Both versions of the BERT model can be fine tuned by adjusting the top and final layers in order to effectively carry out a specific NLP task with a smaller amount of task-specific data (49).

3.4.3 DistilBERT Model

DistilBERT model is a distilled version of the original BERT model. DistilBERT was firstly presented in 2019 by Sanh et al.(50), in order to reduce the computational and resources demands of the original BERT model. The researchers of DistilBERT in order to reduce the computational demands of BERT model, made use of a technique known as Knowledge Distillation that was introduced by Hinton et al. (51). The technique generates a much smaller student model that is trained to replicate the behavior of the larger teacher model. The student model is trained is way that aligns its output to the output of the teacher model. By this procedure the knowledge of the teacher model is transferred to the student model. With this procedure, DistilBERT is able to achieving BERT's 95% performance, while it has

only 40% of BERT's size and 60% more speed (50).

DistilBERT follows the same architectural design of BERT model with some small changes in the parameters. The model is constructed with a number of transformer block that are identical to each other. The transformer blocks include the multi head attention mechanism and the feed forward network, just like the BERT model. In contrast to BERT and RoBERTa, DistilBERT has only one variation, namely DistilBERT base. The DistilBERT base model consists of 6 transformer blocks instead of 12 in BERT model, 768 hidden size layers, 12 attention heads and 66 million parameters instead 110 million parameters in BERT model (50). In the same manner with BERT model, DistilBERT can be fine tuned by adjusting the top and final layers in order to effectively carry out a specific NLP task with a smaller amount of task-specific data.

3.4.4 BERTopic Model

BERTopic is a topic modeling methodology that was introduced in 2021 by Grootendorst (52). The main difference of this approach to traditional topic modeling methodologies is that BERTopic makes uses the capabilities of the transformer based model BERT that was has been already analyzed in this study in order to understand and capture the semantic meaning not only of words but also of documents.

More specifically, BERTopic model has 4 distinct phases. The first phase makes use of the Sentence-BERT model that Reimers and Gurevych proposed in 2019 (53), in order to convert the textual input into word embeddings. These embeddings are able to effectively understand and capture the semantic meaning of the input data.

The second phase is meant for dimensionality reduction. For this purpose BERTopic utilizes a recent approach of dimension reduction called Uniform Manifold Approximation and Projection (UMAP) presented by Maclnnes et al (54). in 2018. Initially, UMAP generates a graph representation of the embedded data in the high dimensional space. After that it generates a graph in low dimensional space and tries to optimize it in order to make it as similar as possible in structure with the high dimensional representation. With this method, the final generated graph is able to maintain all the needed information of the data but with lower dimensions and therefore allow the model to perform a more efficient clustering in the next steps.

The third phase of BERTopic, clusters the documents in the lower dimensional space. For this purpose, the density based clustering algorithm, namely Hierarchical Density-Based

Spatial Clustering of Applications with Noise (HDBSCAN) (9), is employed. HDBSCAN initially transforms the space based on the density of the data points and subsequently clusters data points with single linkage. After clustering, HDBSCAN transforms clusters into cluster tree in order to extract stable clusters.

The final stage of BERTopic utilizes the Class-based Term Frequency-Inverse Document Frequency (c-TF-IDF) in order to determine which words represent more accurate each topic by evaluating how relevant is a specific word to each one of the clusters that were generated in the previous phase. The words with the highest c-TF-IDF scores are selected as the ones that are considered the representative terms of each cluster of topic.

4 DataSet and Exploratory Data Analysis

4.1 Data Description

The dataset that was utilized in the present study is a set of 30,000 tweets referring to ChatGPT. The aforementioned tweets were scraped via Twitter's application programming interface (API) by targeting three relevant hashtags, namely #chatgpt, #chatgptplus, and #gpt4. The decision to gather tweets with the #gpt4 hashtag was made due to the fact that OpenAI decide to give early access to the new language model to all Plus members of ChatGPT. The initial dataset prior the data cleaning procedure included over 45,000 tweets. The data collection was performed from March till April 2023. The time period that the dataset covers span from 30 November 2022 to 8 April 2023, in order to include a total of 130 different days. The time frame of the dataset includes all the significant event of ChatGPT's life spam so far, from the beginning and launch of ChatGPT till the introduction of the Plus subscription and also the introduction of the GPT-4 language model. Moreover, after the tweets were obtained, by using the "Twitter-roBERTa-base for Sentiment Analysis" (55) model they were annotated based on their overall sentiment, into one of the three following categories: "positive," "negative," or "neutral" sentiment. To ensure the preservation of data integrity, only tweets with a probability of belonging to one of the sentiment classes exceeding 80% were chosen. After removing duplicate posts, the size of the final dataset reduced to the previously mentioned number of 30,000 tweets. Finally, each tweet was reprocessed in order to remove any user mentions, hashtags, emojis and last every possible unknown character.

4.2 Exploratory Data Analysis

To begin the exploratory data analysis, the daily volume of tweets was taken into consideration. The findings, as shown in Figure 3, indicate a pattern where the amount of tweets seems to be impacted by some specific events. The amount of the daily tweets, has three important spikes at three important dates. These important date are the launch of the ChatGPT platform in late November 2022, the introduction of the ChatGPT Plus subscription service at the start of February 2023, and also the launch of the GPT-4 model in mid-March 2023. This suggest a possible significant impact that these events may have had in the public sentiment on Twitter towards OpenAI's language models and online platforms.

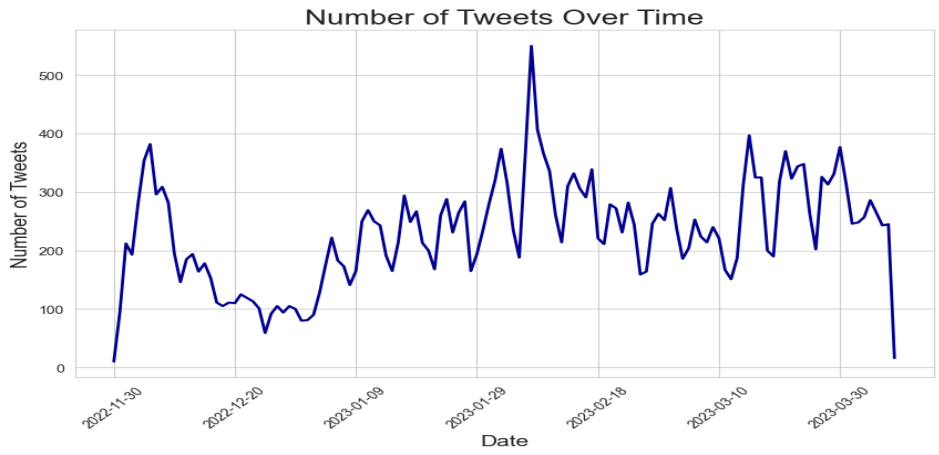

Figure 3: Number of Tweets Over Time

Additionally, an analysis has been conducted on the diverse tweet lengths utilized by users while mentioning ChatGPT. Apparently, the tweets have a significant variety in the tweet length, with the minimum length of a tweet being 13 characters and the maximum tweet length being 305 characters. On average, tweets were approximately 148 characters in length. A more careful review of the distribution of the tweet length as seen in Figure 4 has revealed a distinct bimodal pattern with two prominent peaks around the 150 and 250 character marks, respectively. This pattern suggests that individuals tend to prefer specific character counts when expressing their opinions about ChatGPT and OpenAI.

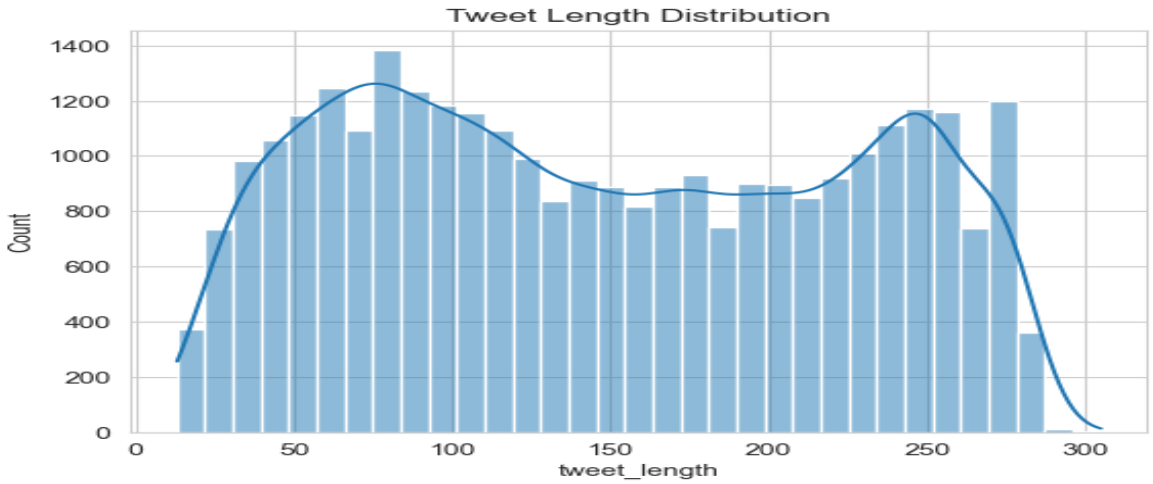

Figure 4: Tweets Length Distribution

The dataset's user verification status provides a deeper understanding for the characteristics of the users that used the aforementioned hashtags. A significant proportion of users, precisely 96.7% were unverified, while a considerably lower proportion of them, precisely 3.3%

were verified. This suggest that the talk about ChatGPT was mainly carried out by ordinary Twitter users, while the verified ones, who in most cases are public figures, made a significantly smaller contribution.

Tweet's sentiment status in Figure 5 demonstrated that the tweets have a balanced combination of positive and neutral sentiment, with a really small fraction of negative ones. The predominant sentiment of the tweets was neutrality, with a percentage of 47.7% of the total tweets. The percentage of the tweets that relate to positive sentiment was only slightly lower, totaling in 42.2%. On the other hand, the tweets with a negative sentiment account only for the 10.1% of the total tweets.

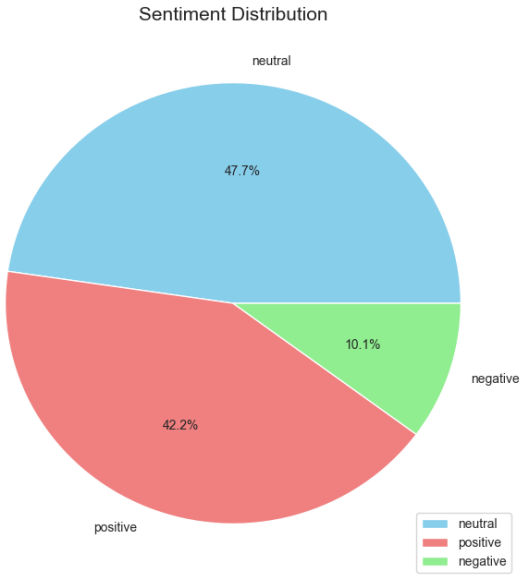

Figure 5: Sentiment Distribution

By analyzing the user verification status and their sentiment status, it became clear that both verified and non verified users have similar patterns. Yet, there was a slight difference in the percentage of neutral sentiment among verified users, as shown in Figure 6

Figure 6: Sentiment Distribution Per User Type

5 Sentiment Over Time Analysis

5.1 Rolling Averages Analysis

To address the initial research question of this study, it is vital to recognise and understand the sentiment trends and shifts that occurred as a result of the transition to a freemium platform with the launch of ChatGPT Plus subscription, and thus to determine if and how strong was it's impact on the general public sentiment. An analysis of the rolling averages of sentiment was used in order to provide an initial overview of the sentiment trends and as a consequence some initial insights about the overall user sentiment before and after the launch of the Plus subscription. The fluctuations in the sentiment can be seen in Figure 7.

The introduction of ChatGPT Plus subscription service seems to have resulted in an increase in the neutral sentiments, along with a small increase in negative sentiments. This increase in the percentage of users with a neutral attitude implies that there is a possibility of some skepticism within the users regarding the platform change. This trend could be interpreted as users hold a more cautious approach with the changes that OpenAI introduced to the platform and avoid to express their sentiment until they have had the opportunity to assess the newly changed platform and decide if the change is valuable for them or not.

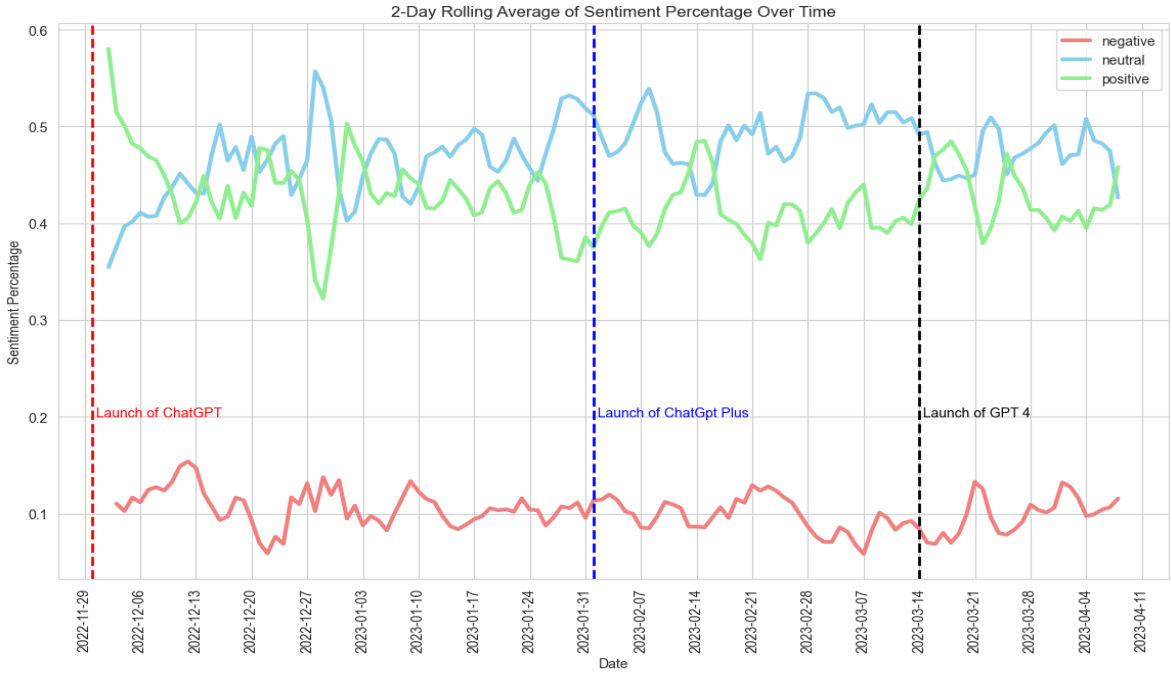

Figure 7: 2-Day Rolling Average of Sentiment Percentage Over Time

On the other hand, the small rise in negative emotions indicates that there is a small group

of users who might have difficulty to accept the implementation of the subscription based model. However, the increase in negative sentiments was not big enough to be considered significant and important. Such a negative response may have been triggered by concerns about for how long will the platform be accessible for free users and the overall cost of the platform.

Moreover, the introduction of the GPT-4 model has resulted in an increase in positive sentiments and an equal decrease in neutral sentiment. This trend may have been triggered by the fact that users are always interested and excited with significant developments in AI. This rising enthusiasm could be the reflection of users anticipation of using a new language model such as GPT-4.

Furthermore, it is noteworthy to mention that between the dates of December 25, 2022, and the initial days of 2023, there was a notable fluctuation in the number of neutral tweets, characterized by an abrupt increase followed by a subsequent decrease. Similarly, there was a corresponding decline succeeded by an upsurge in the volume of positive tweets during the same time period. During the aforementioned time frame, speculations regarding a potential acquisition of OpenAI by Microsoft were a significant source of concern among users.

While these interpretations based on rolling averages can provide an initial overview of public sentiment over time, they also show a more temporary presentation of the users sentiment. In order to decide if these shifts in users mood has a significant effect and not only temporal, a statistical method will be used in the next section.

5.2 Interrupted Time Series Analysis

A statistical validation technique that can be used to evaluate the results of the rolling averages analysis is the combination of OLS regression and an interrupted time series analysis (ITSA). In this way, the influence and significance of an intervention can be evaluated.

The null hypotheses are the following two:

1. H_0*: There is no effect in user sentiment following the introduction of the Plus subscription.*

2. H_0*: There is no effect in user sentiment following the release of the GPT-4 model.*

Based on Table 3 the results of the OLS regression showed that the effect of the launch of Plus subscription on the public sentiment has a coefficient of -0.0181 and a p-value of 0.209. Since the p-value is higher than the significance level of 0.05, the first null hypothesis of the

first H_0 can not be rejected. This suggest that the sentiment shifts that occurred after the launch of the Plus subscription might not have been statistically significant.

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
Intercept	0.3302	0.009	36.751	0.000	0.312	0.348
Post ChatGPT Plus	-0.0181	0.014	-1.263	0.209	-0.046	0.010
Post GPT-4	0.0103	0.009	1.154	0.251	-0.007	0.028
Interaction Of Both Events	0.0103	0.009	1.154	0.251	-0.007	0.028

Table 3: Interrupted Time Series OLS Results

The corresponding p-value and the coefficient for the sentiment shifts that followed the launch of GPT-4 was 0.251 and 0.0103 respectively. Once again based on the significance level of 0.05 the null hypothesis of the second H_0 can not be rejected. This shows that the sentiment changes that occurred after GPT-4's debut were not statistically significant.

Moreover, the interaction term between both events resulted in a coefficient of 0.0103 and a p-value of 0.251. This result shows that the combined influence of these two events on user sentiment was not statistically significant on a significance level of 0.05. Interestingly, the results for the interaction term are the same as the ones for the Post GPT4 event. This may happened due to the use of binary variables in the model or simply the model is over specified and the interaction term do not add anything more in to the model.

Finally, the combination of rolling averages and ITSA results provided a deep understanding on the user's sentiment. The rolling averages emphasized on the initial shifts of sentiment by the introduction of the Plus subscription and the debut of GPT-4. The fact that these changes were not statistically significant as shown by the ITSA results, suggests that these changes did not play crucial role on public sentiment on the long run and therefore the affect of the introduction of the Plus subscription did not affect the public sentiment, but it rather produced some temporary fluctuations in the overall sentiment among the users.

6 Sentiment Analysis Modelling

In the following part of this study, a broad range of models will be tested and then compared against one another to determine which model is better suited for performing sentiment analysis. The baseline models include the Support Vector Machine (SVM) and Naive Bayes model, both of which are machine learning algorithms, as well as Convolutional Neural Network (CNN), witch is a deep learning model. Furthermore, cutting edge pre trained models such as BERT model, RoBERTa model, DistilBERT model and a combination of DistilBERT with CNN model will be examined.

6.1 Machine Learning Algorithms

6.1.1 Support Vector Machine

Starting off with the Machine Learning algorithms, this study presents a sentiment classification model using the SVM algorithm in combination with the One-vs-Rest (OvR) approach. Due to the fact that the tweets express three distinct sentiments the SVM model alone would not be appropriate as it does not support multi class classification by default. The proposed model with the addition of the OvR approach can distinguish the sentiment between our three targeted classed as one classifier is used for each class against all other classes.

In addition to the initial pre processing of the data set, the data is transformed into numerical features using the Term Frequency- Inverse Document Frequency (TF-IDF) vectorization approach. After the vectorization process, the data set was divided into two subsets with a 80:20 ratio for training and testing set.

During the hyper parameter optimization phase of the SVM model, the penalty parameter C that controls the trade off between increasing the margin of separation between sentiment classes and reducing the classification error was determined to be 100 after using a grid search and 5 fold cross validation.

	Precision	Recall	F1-Score	Support
Negative	0.77	0.44	0.56	632
Neutral	0.77	0.86	0.81	2841
Positive	0.82	0.81	0.81	2527
Accuracy			0.79	6000
Macro Avg	0.79	0.70	0.73	6000
Weighted Avg	0.79	0.79	0.79	6000

Table 4: SVM Classification Report

Moreover, as seen in Table 4 and Figure 8, the evaluation of the model on the separate test set showed that the model performed with an overall accuracy of 79%. For the negative, neutral, and positive sentiment classes, the model achieved precision values of 0.77, 0.77, and 0.82. In addition, the model also has recall values of 0.44, 0.86, and 0.81 for negative, neutral and positive classes. This suggests that although the model performs relatively well at identifying the neutral and positive tweets, it has a difficulty at classifying tweets with a negative sentiment. This could be closely associated with the class imbalances of the dataset.

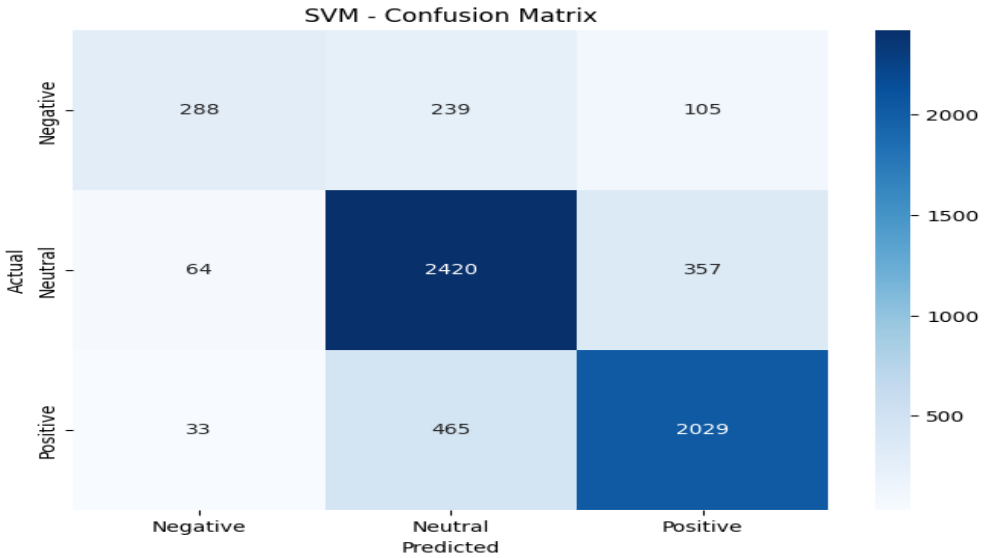

Figure 8: SVM Confusion Matrix

In conclusion, the suggested SVM model with the use of the OvR method performs well in the categorization of sentiment. However, the model is unable to accurately classify all classes. Negative tweets are classified with a significantly lower accuracy than other classifications.

6.1.2 Naive Bayes

Next, a Naive Bayes model was used. The same procedures as for the prior SVM model were carried out in terms of preprocessing and data split. The Multinomial Naive Bayes classifier was trained on the training data and afterwards evaluated on the separate test set.

The overall accuracy was found to be around 73% as shown in Table 5.

	Precision	**Recall**	**F1-score**	**Support**
Negative	0.00	0.00	0.00	632
Neutral	0.74	0.80	0.77	2841
Positive	0.72	0.83	0.77	2527
Accuracy			0.73	6000
Macro avg	0.49	0.54	0.51	6000
Weighted avg	0.65	0.73	0.69	6000

Table 5: Naive Bayes Classification Report

Based on the metrics for each of the three classes, it is evident that the model performed relatively well for neutral and positive tweets. The precision, recall and F1 score values for both of these two classes suggest that the model has the ability to detect with fairly good accuracy tweets with neutral or positive sentiment.On the other hand, the model displayed a lack of ability to correctly classify any tweet with a negative sentiment.The negative sentiment class has zero values for all precision, recall and F1 score. The visual representation of this inability can also be observed in Figure 9.

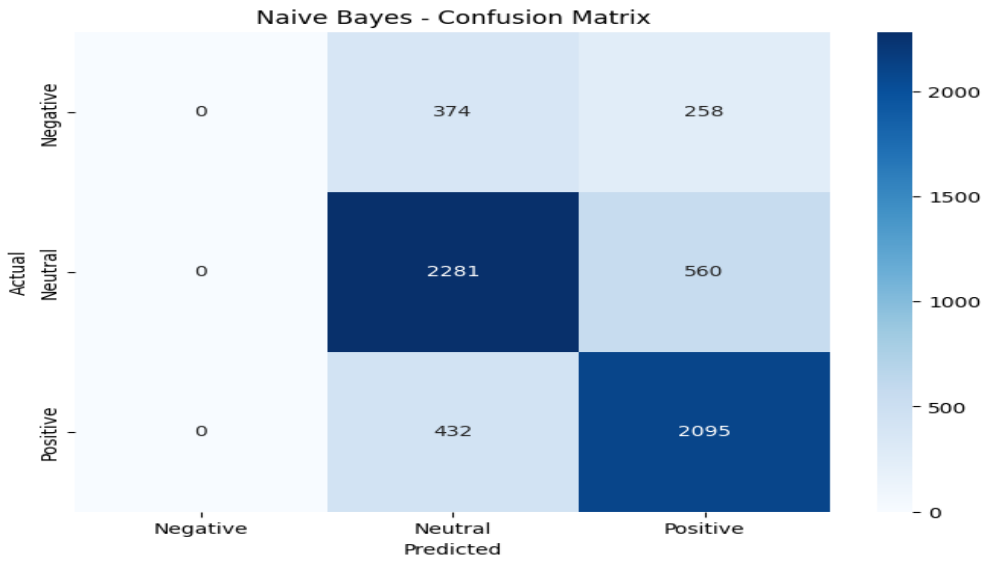

Figure 9: Naive Bayes Confusion Matrix

The observed complete inability of the model to not even detect a single instance with a negative sentiment can be attributed to various factors. One potential issue could be the imbalance of classes in the dataset as already stated that the negative class accounts for only 10% of the total number of tweets.

6.2 Deep Learning

6.2.1 Convolutional Neural Network

Moving on to the Deep Learning realm, a Convolutional Neural Network model was utilized for the sentiment analysis task. The architecture of the proposed model is visually presented in Figure 10. The Keras API was utilised to construct the model with a sequential architecture. One of the key layers of the proposed model is an embedding layer that produces a compact representation of each word and the corresponding meaning of it. Following that, the processed data is inputted into a pair of Conv1D layers, in which each layer consists of 128 and 64 neurons, respectively. In order to reduce the potential issue of over-fitting, an L2 regularisation term was introduced at the two convolutional layers.

Following the convolutional layers, the MaxPooling1D and the GlobalMaxPooling1D pooling layers are used. The use of the pooling layers is to reduce the spatial dimensions of each sequence, thereby reducing computational demands. Furthermore, this reduction in complexity enables the model to better generalise to unseen data.

Next, there is a Dense layer and a Dropout layer that randomly decides to ignore the 30%

of the inputs during each iteration of the training process in order to reduce the possibility of over-fitting of the model. The final layer of the proposed model is a Dense layer with three nodes, with each node corresponding to each one of the three sentiment classes of our multi class classification. The dataset was divided into three subsets. Randomly 60% of the data was allocated to the training data set, 20% to the validation set and the remaining 20% to the testing set.

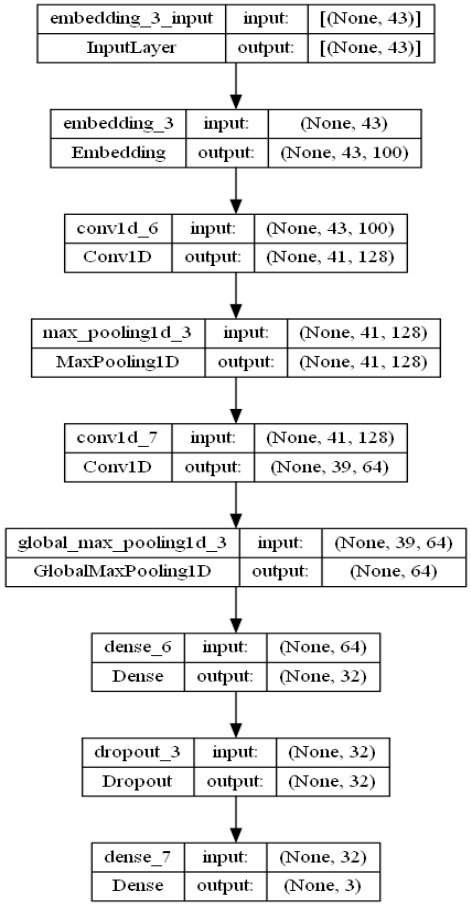

Figure 10: CNN Model Architecture

During the training phase, the model's training accuracy kept improving gradually, resulting in a max of 96.90% at the 10th epoch. However, an early stopping mechanism was implemented in order to prevent the issue of overfitting. At the 10th epoch, this mechanism was activated as the validation loss stopped improving, and the optimal training values were restored. The epoch that produced the optimal training values was the fifth. The learning history of the model is presented in Figure 11.

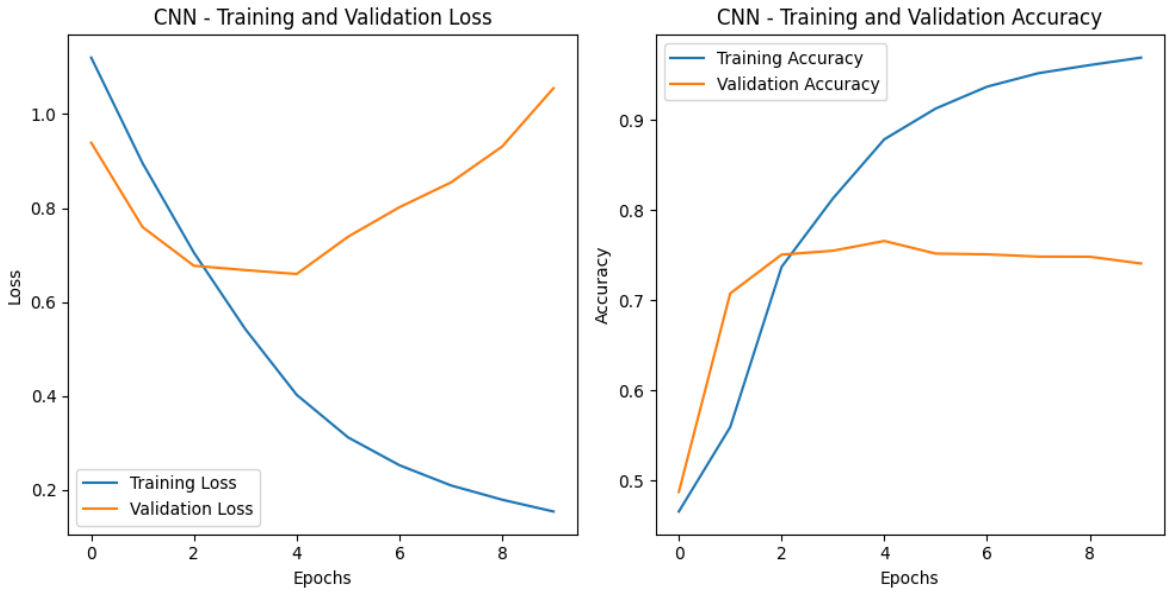

Figure 11: CNN Learning History

The classification report in Table 6 and the confusion matrix in Figure 12 suggest that the model achieved an overall accuracy of 76.82% on the test set. The model showed adequate results at classifying tweets with neutral and positive sentiment, while the metrics were relatively lower for the negative category.

	Precision	**Recall**	**F1-Score**	**Support**
Negative	0.59	0.49	0.53	632
Neutral	0.77	0.82	0.79	2841
Positive	0.80	0.78	0.79	2527
Accuracy			0.77	6000
Macro Avg	0.72	0.70	0.70	6000
Weighted Avg	0.77	0.77	0.77	6000

Table 6: CNN Classification Report

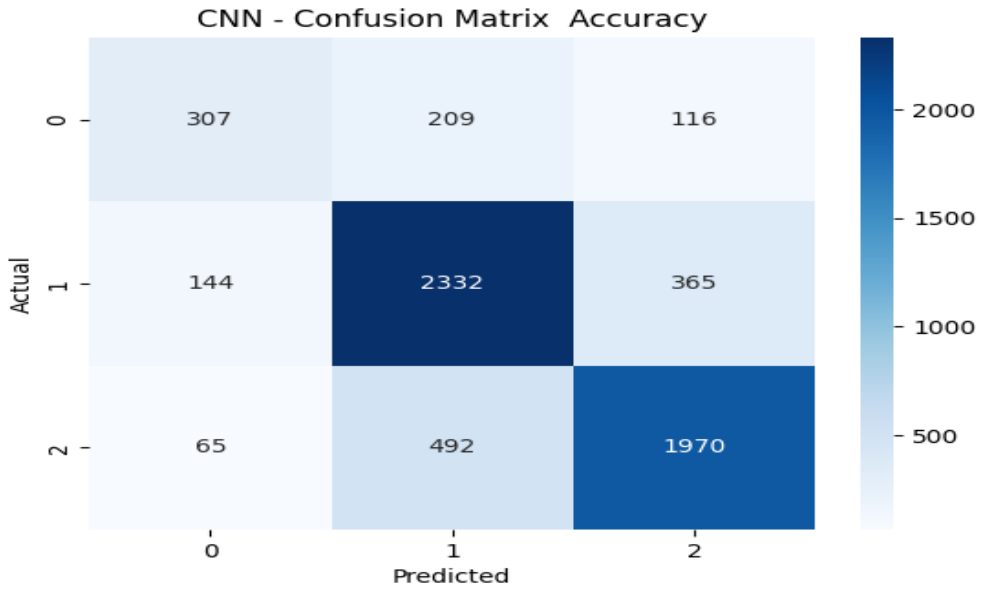

Figure 12: CNN Confusion Matrix

In summary, the proposed convolutional neural network demonstrated strong performance for sentiment analysis classification, particularly in accurately classifying neutral and positive tweets. The performance on negative tweets was lower than the other two classes but much better than the ones of Naive Bayes.

6.3 Large Language Models

6.3.1 BERT Model

Moving on to pre trained deep learning models, a BERT-based model for sentiment analysis was employed.

The proposed model used BERT model's underlying architecture, in this case the "bert-base-uncased" (46), witch allows it to comprehend the semantics of the input text by giving a 768-dimensional vector representation. Moreover, additional layers were incorporated into the architecture in order to help the model acquire knowledge from the provided data. After the BERT model layer, a linear layer with 768 neurons was introduced to transform the input data to an appropriate form for the following layers. A Dropout layer was included with a rate of 0.3 to prevent over fitting. The final layer of the model is again a linear layer with 3 possible outputs in order to match the three sentiment classes of the data.

The model was trained for five epochs. The training and validation accuracy of the model had a progressive improvement after each epoch as seen in Figure 13. During the first epoch,

the model achieved a training accuracy of 79.15% and a validation accuracy of 80.02%. By the second epoch, the training and validation accuracy had increased to 88.56% and 83.85%, respectively. The improvement continued until the third epoch, at which point the model obtained a training accuracy of 93.81% and a validation accuracy of 83.63%. However, after observing an increase in validation loss during the third epoch,the early stopping rule of the model was triggered in order to not let the model get too accustom to training data set and therefore start overfitting. As a result, the model's optimal weights were reset to those that were obtained during the second epoch in which the lowest validation loss has been observed.

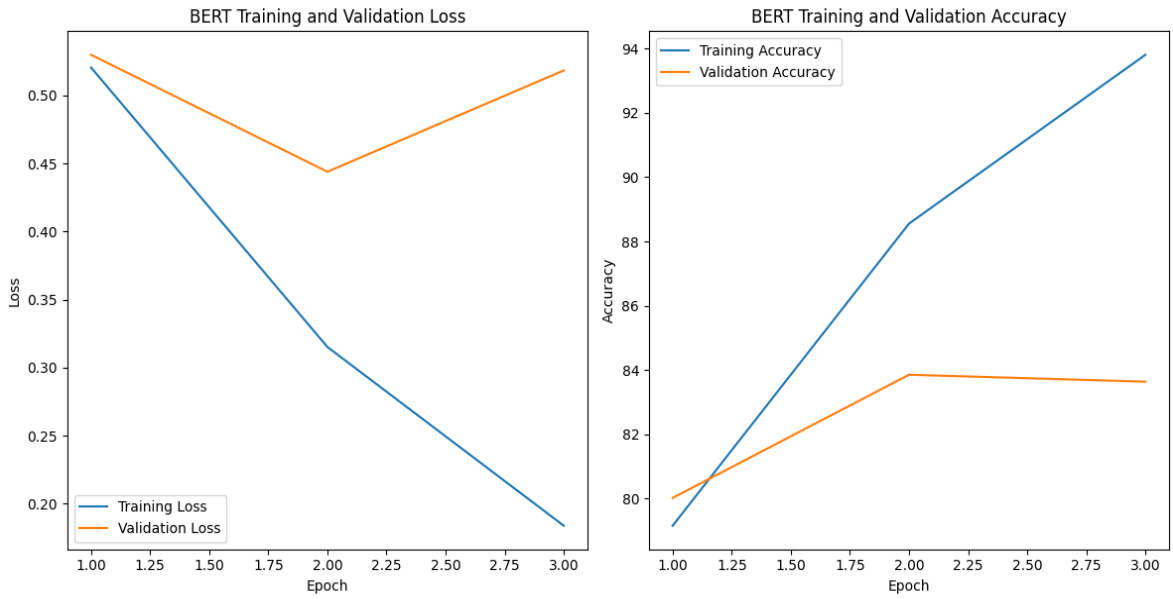

Figure 13: CNN Confusion Matrix

After the training phase, the model was tested on the separate test set to assess its performance. The results of the testing showed that the model has a great ability to classify the tweets into their corresponding sentiment class with an overall accuracy of 84% as can be seen by the classification report in Table 7.

The precision rates for positive and neutral sentiments were observed to be significantly high, with values of 0.86 and 0.83, respectively. However, the precision rate for the negative sentiment class was somewhat lower. Furthermore, the model has good recall rates for the three sentiment classes , specifically 0.85 for positive emotions and 0.87 for neutral emotions.

	Precision	Recall	F1-Score	Support
Negative	0.81	0.65	0.72	632
Neutral	0.83	0.87	0.85	2841
Positive	0.86	0.85	0.86	2527
Accuracy			0.84	6000
Macro Avg	0.83	0.79	0.79	6000
Weighted Avg	0.84	0.84	0.84	6000

Table 7: BERT Classification Report

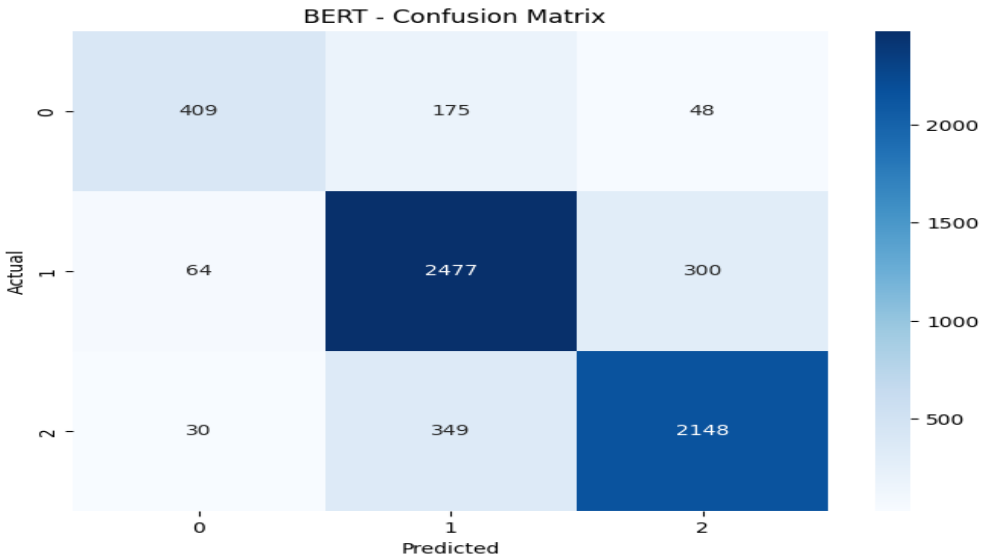

Figure 14: CNN Confusion Matrix

While the model performed well in recognizing positive and neutral attitudes, it performed somewhat worse in categorising negative sentiments, with a precision of 0.81 and a recall of 0.65. This can be made more understandable through the model's confusion matrix in Figure 14. But overall, the model showed a considerable improvement compared to the previously examined models.

6.3.2 RoBERTa Model

To continue,the RoBERTa model, which is an advanced variation of the transformer based deep learning BERT model,was employed. The proposed model used RoBERTa model's underlying architecture, in this case the "roberta-base" (49), which allows it to comprehend the semantics of the input text by giving a 768-dimensional vector representation in the

same way as the previous BERT model. Moreover, three additional layers were implemented into the architecture. Following the RoBERTa layer, a linear layer with 768 neurons was introduced. To continue a Rectified Linear Unit (ReLU) activation function and a Dropout layer were used with a rate of 0.4 to prevent over fitting. The final layer of the model was again a linear layer with a 3 dimensional output in order to match the three distinct sentiment classes of the data.

The model was trained over three epochs. The model achieved a training accuracy of approximately 78.44% and a validation accuracy of approximately 75.91% during the first epoch.With the second epoch the model had a significant performance increase.

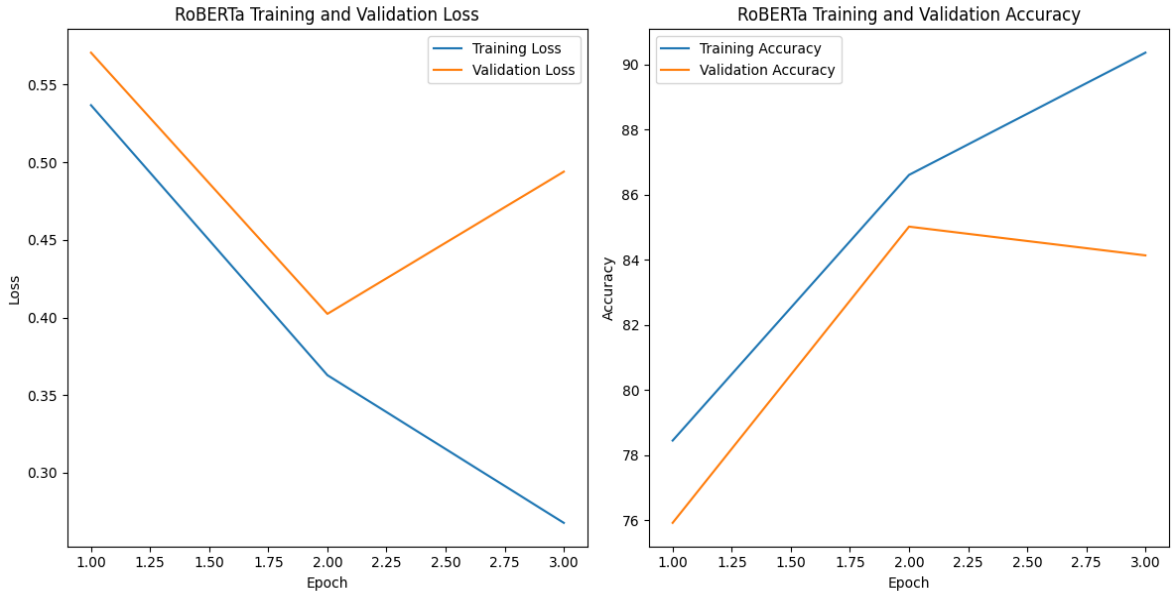

Figure 15: RoBERTa Learning History

The training accuracy increased to 86.60% and the validation accuracy increased to 85.01%. During the third epoch the training accuracy continued to increase to reach 90.36%, however the validation accuracy decreased to 84.13%. This decrease in validation accuracy may indicates the possibility of overfitting. The learning history of the model can be seen in Figure 15.

Although there were some slight indications of overfitting, the model showed exceptional performance on the test set, as shown in the classification report of Table 8. Upon testing the model in the test the model achieved an overall accuracy of 84.5% as Figure 16 shows. Furthermore, the model showed a balanced performance across all three sentiment classes.

	Precision	Recall	F1-Score	Support
Negative	0.83	0.65	0.73	632
Neutral	0.86	0.83	0.85	2841
Positive	0.83	0.91	0.87	2527
Accuracy			0.85	6000
Macro Avg	0.84	0.80	0.82	6000
Weighted Avg	0.85	0.85	0.84	6000

Table 8: RoBERTa Classification Report

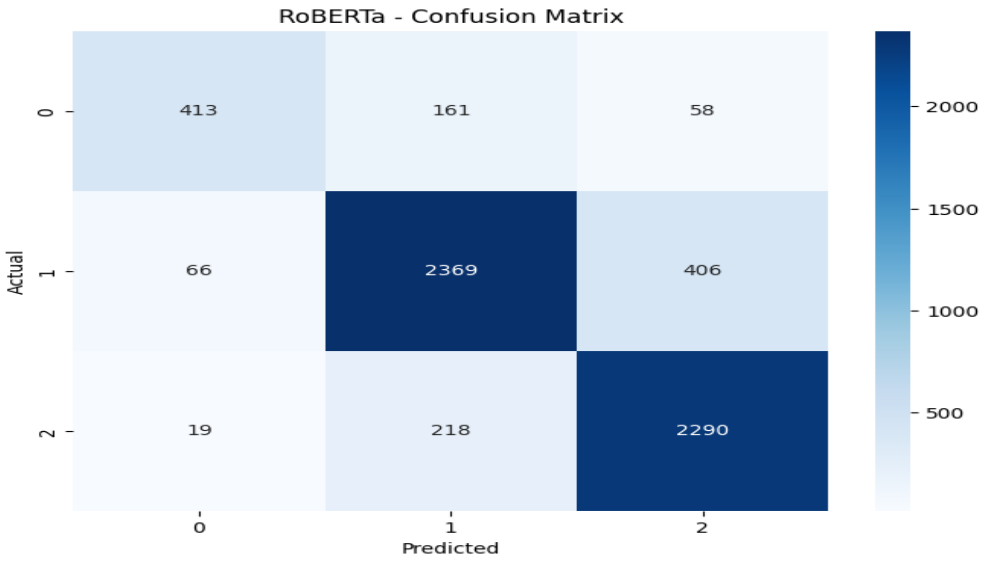

Figure 16: RoBERTa Confusion Matrix

6.3.3 DistilBERT MOdel

The next model of this study is a DistilBERT model.The proposed model was built on DistilBERT's architecture, in this case the "distilbert-base-uncased" (50). Three additional layers were added to the original DistilBERT model, namely a pre-classification linear layer, a final classification linear layer and a dropout layer. These additional linear layers were incorporated in order to adapt the model for the multi class classification. Moreover, the dropout layer was included in order to prevent over fitting.

The training of the model was done over a span of three epochs, as shown in Figure 17. Throughout the training process, a steady increase was observed on the training set. To be more specific, the accuracy began at 77.91% in the first epoch,increased to 86.75% in the second, and finally reached 91.58% in the third.

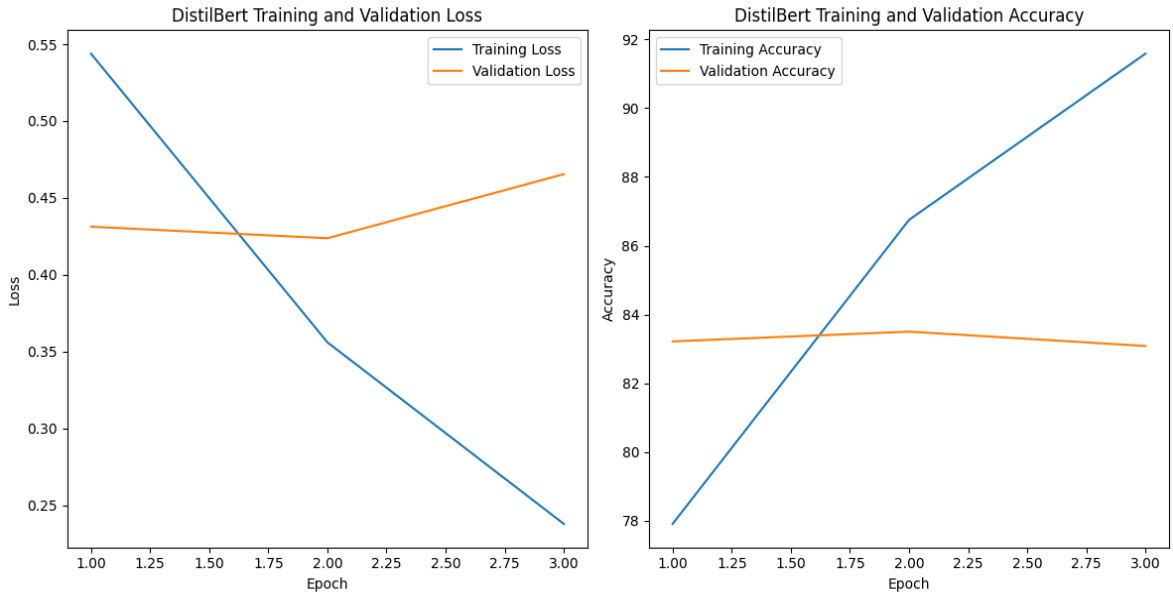

Figure 17: DistilBERT Learning History

Throughout each of the three epochs, the model's accuracy on the validation set remained rather stable at approximately 83%. Upon evaluation on the test set, the model showed good performance and ability to generalize on unseen data,achieving an overall accuracy of 83% which is exactly the same as the model's accuracy on validation set during the training process. The classification report presented in Table 9 shows that the model performed reasonably well across all the three sentiment classes.

	Precision	**Recall**	**F1-Score**	**Support**
Negative	0.82	0.60	0.69	632
Neutral	0.83	0.85	0.84	2841
Positive	0.84	0.86	0.85	2527
Accuracy			0.83	6000
Macro Avg	0.83	0.77	0.79	6000
Weighted Avg	0.83	0.83	0.83	6000

Table 9: DistilBERT Classification Report

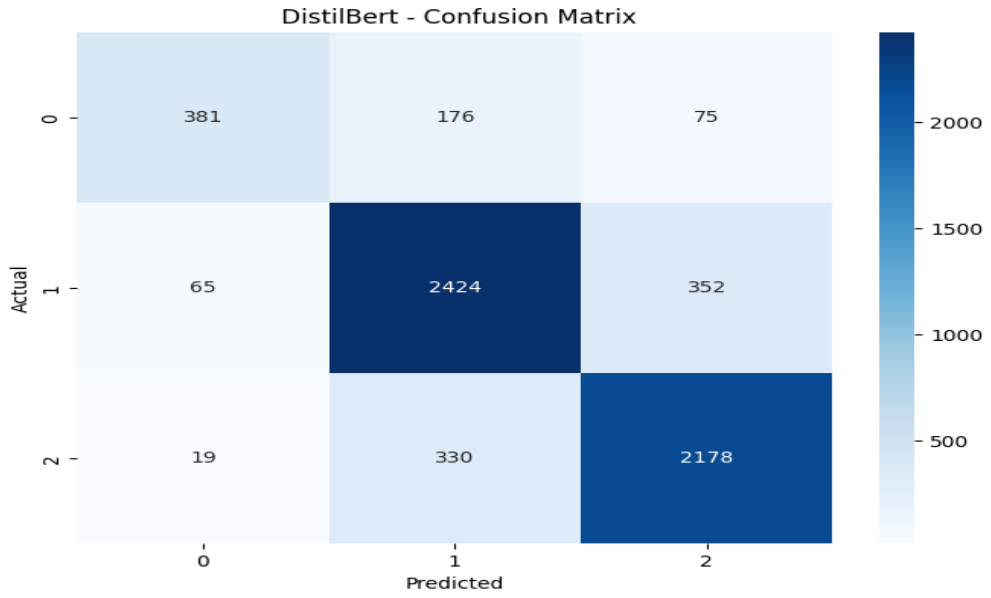

Figure 18: DistilBERT Confusion Matrix

In summary, the proposed DistilBERT model showed a strong performance in terms of generalization to previously unseen data as the confusion matrix of Figure 18 shows,despite the fact that there were some indications of overfitting during the training process.

6.3.4 DistilBERT And CNN Model

Finally, the last model that was employed for this study was a combination of a DistilBert model with a CNN model. The proposed model used the DistilBert's architecture, namely the 'distilbert-base-uncased' (50) as the base layers in order to convert raw input text into meaningful representations. The following layers were a linear layer, a dropout layer, and another linear layer. The proposed model also incorporates a 1-dimensional convolutional layer and a max pooling layer that come in directly after the DistilBert layers.

The training process of the model was done in a span of two epochs due to the fact that when the epochs were extended to three, a significant increase in validation loss was observed. The training accuracy showed an increase from from 78.17% to 86.92%, while the validation accuracy demonstrated a decrease from 82.83% to 83.7% as can be seen in Figure 19. The model showed solid performance on the test set, achieving an overall accuracy rate of 83.6%.

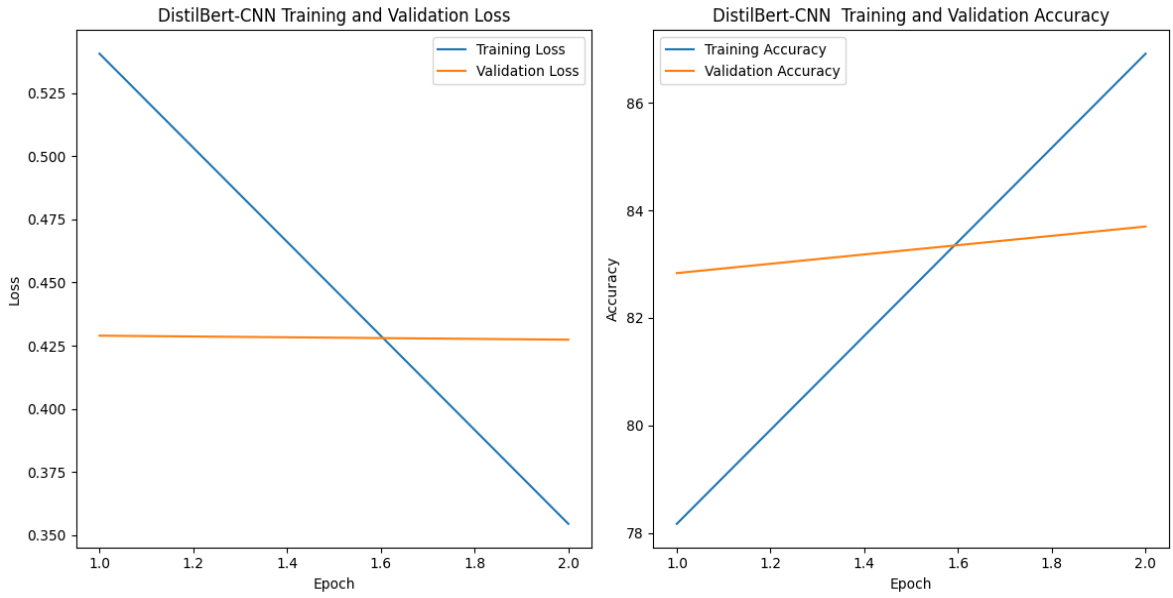

Figure 19: DistilBERT & CNN Learning History

Based on the classification report of Table 10, the performance metrics were not consistent across all three different classes. The accuracy of negative instances was 75%, while for neutral and positive classes the accuracy was 84%. The aforementioned observation suggests that the model encounters some challenges in identifying negative tweets in contrast to the remaining two sentiment classes. However, the difference in the accuracy of the model across the various classes is not substantial.

	Precision	**Recall**	**F1-Score**	**Support**
Negative	0.75	0.69	0.72	632
Neutral	0.84	0.85	0.85	2841
Positive	0.84	0.86	0.85	2527
Accuracy			0.84	6000
Macro Avg	0.81	0.80	0.81	6000
Weighted Avg	0.84	0.84	0.84	6000

Table 10: DistilBERT & CNN Classification Report

The difficulties associated with distinguishing negative tweets from other categories are evident also in Figure 20.

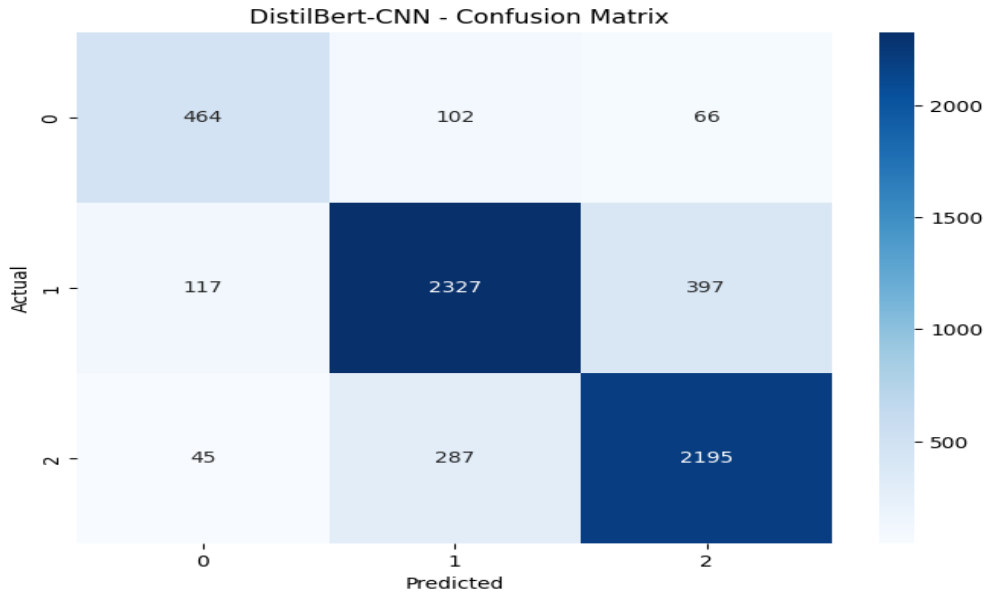

Figure 20: DistilBERT & CNN Confusion Matrix

To summarize, the proposed combination of DistilBert and CNN presents as a valuable solution for sentiment analysis. However, there are certain inconsistencies in its performance with respect to specific classes.

6.4 Models Comparison

To begin the comparison of the models previously discussed in the study, Table 11 shows that the conventional machine learning models, specifically the SVM model and the Naive Bayes model, had a notable difference in their individual performances. The SVM model with an overall accuracy of 79% outperformed the Naive Bayes model not only in terms of accuracy but in every single other metric.

Model	**Precision**	**Recall**	**F1-score**	**Accuracy**	**False Rate**
SVM	0.79	0.79	0.79	0.79	0.21
Naive Bayes	0.65	0.73	0.69	0.73	0.27
CNN	0.77	0.77	0.77	0.77	0.23
BERT	0.84	0.84	0.84	0.84	0.16
RoBERTa	0.85	0.85	0.84	0.85	0.15
DistilBERT	0.83	0.83	0.83	0.83	0.17
DistilBERT + CNN	0.84	0.84	0.84	0.84	0.16

Table 11: Model Comparison

On the other hand, the performance of the Naive Bayes model could be described at best as moderate. The assumption that the Naive Bayes model makes that all the features are independent, could be one of the most important reasons for the poor performance, as the model tries to over simplify complex textual data. The performance of the Naive Bayes model was found to be way inferior to that of the SVM model and all the other models in comparison.

To continue, the CNN model outperformed the Naive Bayes model but its performance was inferior to the performance of the SVM model. However the difference between the two of them was not big enough to be considered significant.

The use of transformer-based models, specifically BERT, RoBERTa, and DistilBERT, has resulted in notable improvements in performance. The BERT model demonstrated superior performance compared to its prior model,as evidenced by its precision, recall, and F1-score and overall accuracy. The RoBERTa model outperformed all other models as it achieved the highest accuracy at the level of around 85% and as a consequence the lowest misclassification error rate of around 15%. It is worth noting that the proposed DistilBERT model demonstrated comparable performance with all the other transformed based models even if the model is a reduced variation of the larger BERT model. This indicates that the model can be a good choice in situations where computational resources may be limited. Finally, the proposed combination of DistilBERT and a CNN model, despite demanding more computational resources than the other models, produced results that were comparable to those of the DistilBERT model. Although the model provides good results, the expense of the increased complexity and computational demands cannot be rationalised, as it required almost twice the computational time of the DistilBERT model.

In conclusion, it can be stated that the RoBERTa model is the superior option among all the other models, when the highest level of precision and accuracy is the number one priority. However, when considering additional factors and variables, such as computational resources, a model like DistilBERT could be a good option, as it strikes a balance between adequate performance and decreased computational demands.Moreover as a final takeaway, the state of art transformer based models outperform both machine learning algorithms and simple deep learning model for sentiment analysis.

6.5 Factors That Influence Public Sentiment

The aim of this section is to determine the factors or topics that have an influence on the users' sentiment towards OpenAI and ChatGPT. In order to identify the aforementioned

topics, a BERTopic model will be applied to the dataset. By analyzing the subjects provided by BERTopic, a deep understanding of the factors that influence public's sentiment regarding OpenAI and ChatGPT can be attained. This, in consequence, could help in possible new strategic decisions.

In the initial application of the BERTopic model on the dataset without any tuning, the model successfully identified and grouped the tweets into more than 300 distinct topics. Upon examination of these given topics, it was observed that several topics had a significant degree of overlap with one another. Following many iterations of trial and error, a decision was made to set a maximum limit of 50 topics. Thus, the model effectively grouped all subtopics into 50 primary topics. The 20 most important topics can be seen in Figure 21

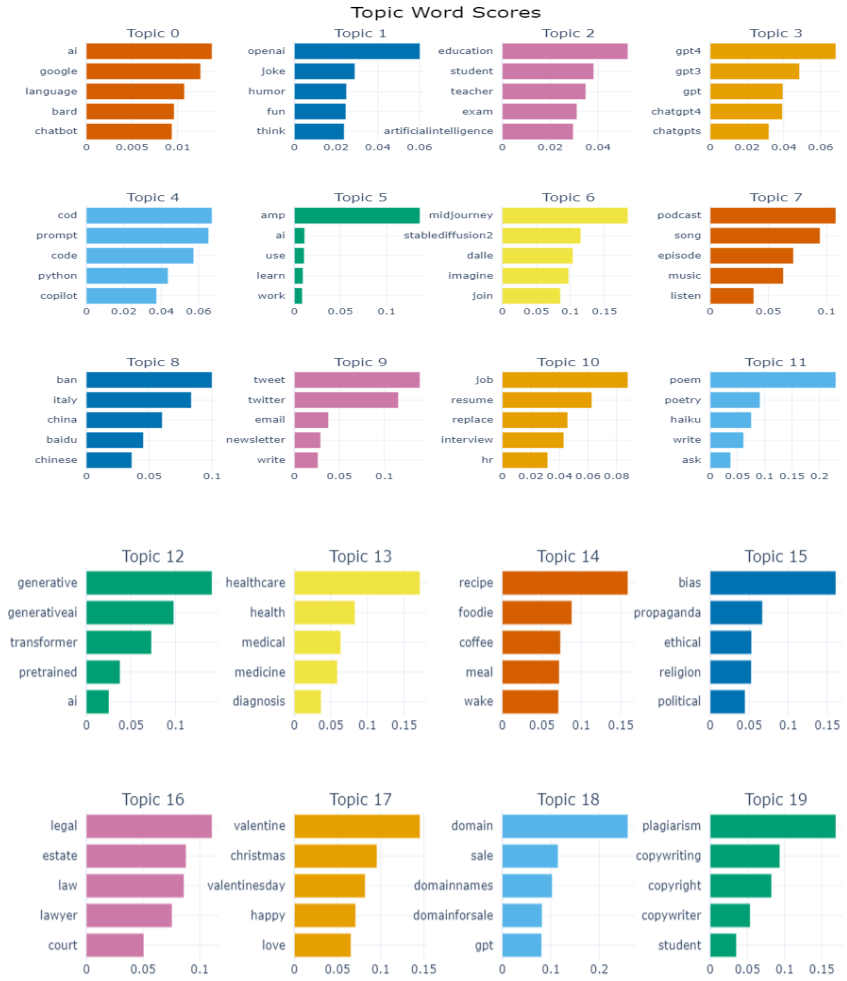

Figure 21: Top 20 Topics

The interpretation of these Top 20 topics could be inferred as follows in the description

column of both Table 12 and Table 13. Furthermore,the tables show the percentage of sentiment in the tweets related to each respective topic. The results indicate that the sentiment regarding ChatGPT is highly valued for its contribution to special events, as the sentiment of this topic was positive by 71.62%. On the other hand, the topic "Digital Marketing" had the most neutral tweets as the neutral emotion holds a 70.76% of the total sentiment of the topic. Finally, the topic "Ethical Considerations in the Deployment" is the topic with the most negative sentiments towards ChatGPT and OpenAI with 37.6% of the overall sentiment.

Topic	**Description**	**Negative %**	**Neutral %**	**Positive %**
Comparison to AI Giants	ChatGPT compared with other AI and tech companies like Google and big language models like Bart	8.71	48.93	42.36
Humour Generation and Understanding	ChatGPT's ability to generate humour and understand jokes	15.91	43.25	40.83
Education	The role of ChatGPT in education and its potential impacts	9.41	49.85	40.74
Evolution of ChatGPT	The evolution of ChatGPT and the various versions of GPT model	8.04	49.30	42.66
Coding and Programming	The contribution of ChatGPT in coding and programming	9.04	43.05	47.92
Practical Everyday Applications	Practical everyday use & applications of ChatGPT	6.93	38.21	54.85
Content Generation Tools	The use of ChatGPT together with other AI generation tools and platforms like Midjourney, Dalle and StableDifussion2	2.34	69.23	28.43
Media and Entertainment	The influence of ChatGPT in media and entertainment tools and sites	6.24	47.26	46.50
Regulation and Policy Implications	Regulation and policy implications for ChatGPT	31.28	52.96	15.76
Generating emails and tweets	ChatGPT and automation of tasks like writing emails or even tweets	10.85	46.91	42.24

Table 12: Sentiment Analysis by Topic with Descriptions (Topics 0-9)

Topic	**Description**	**Negative %**	**Neutral %**	**Positive %**
Implications for Job Market	Impact on job market and HR practices	9.73	48.10	42.17
Literacy Creativity	Explores the literacy creativity shown by ChatGPT	8.15	47.90	43.95
Underlying Technologies	Examination of the technologies that ChatGPT platform is built on, such as transformer based architectures and generative AI	2.71	56.90	40.39
Healthcare	Impact of ChatGPT in Healthcare, including possible help for healthcare practitioners in tasks like diagnosis	4.07	50.23	45.70
Recipes Generation	The ability of ChatGPT to generate food and beverage recipes	14.75	43.17	42.08
Ethical Considerations	Ethical issues that arise with technologies like ChatGPT and considerations for its use	**37.60**	44.96	17.44
Legal Applications	Applicability of ChatGPT in a legal and court context and its potential use by lawyers	9.09	65.78	25.13
Special Occasions	Use of ChatGPT in special occasions like Christmas and Valentines day for gift selection or planning celebrations	5.86	22.52	**71.62**
Digital Marketing	Utilization of ChatGPT in digital marketing and its potential assistance to marketing campaigns	1.17	**70.76**	28.07
Copyright Issues	Discussions about ChatGPT and its relationship with copyright issues and plagiarism	12.73	54.24	33.03

Table 13: Sentiment Analysis by Topic with Descriptions (Topics 10-19)

7 General Discussion

7.1 Key Findings

This study aimed at examining and answering four distinct research questions. Firstly, if adopting a freemium business model affects the public sentiment of users, secondly whether LLM can have a significant improvement on the task of sentiment analysis, thirdly if the combination of an LLM and a deep learning model could potentially outperform other strategies and finally to identify the factors or topics that play significant role in public sentiment towards OpenAI and ChatGPT.

One of the key conclusions of the study is that, the adoption of a freemium business model by an AI platform like ChatGPT did not resulted a sustained impact on the public sentiment of the users and the public in general. Even thought that right after the implementation of the freemium model, there were some initial fluctuations in the sentiment of public, it is important to note that these fluctuations did have statistical significance. This statistical insignificance implies that the transition to the freemium model did not have an impact on public sentiment on the long run. This also implies that users and public have a general flexibility in response to shifts and changes in business models.

In regard to sentiment analysis models, the study showed that there is a significant difference in performance of LLMs compared to all others machine and deep learning models. RoBERTa model demonstrated the best performance against all other models and achieved an accuracy rate of approximately 85%. It is also important to note that DistilBERT model demonstrated also comparable performance while also using less computational resources. The above indicate that LLMs enhance the precision in the task of sentiment analysis compared to more traditional approaches. Another key findings of the study is that the proposed combination of DistilBERT and a CNN model did not improve the classification performance of the DistilBERT model and also did not outperform other LLMs.

Lastly, the study examined the factors and topics that affect the sentiment surrounding OpenAI and ChatGPT. The findings of the study showed that the topics are associated with various subjects like its possible use in various sectors like healthcare, law and digital marketing, among many others. Moreover, the identified topics were linked with a positive sentiment, and others mostly to neutral or negative sentiment. For instance, negative sentiments were associated with regulation and policy implications of ChatGPT, copyright issues like plagiarism, and some ethical considerations. Neutral sentiments were associated with topics referring to

legal applications of ChatGPT, the use of ChatGPT within digital marketing context and combination of it with other content generation tools. On the other hand the positive tweets were focused around the use of ChatGPT during special occasions like Christmas, coding and programming and lastly about practical everyday use of the ChatGPT.

7.2 Managerial Implications

The findings of this study can have some significant implications from a managerial standpoint. First and foremost, the findings hold a significant importance for AI platforms and potentially other online tech platforms that are contemplating a transition from a completely free platform to a freemium business model. In light of these conditions, companies and manager can have the confidence to implement a freemium business model without fearing that such a transition would affect their reputation. However, every company that is considering to make this transition should create and implement appropriate communication strategies that can address effectively the initial and temporary fluctuations in public sentiment. Secondly, the findings from the comparison of sentiment analysis models, highlight the necessity for companies to embrace and adopt more sophisticated models like RoBERTa compared to traditional approaches in order to better understand the sentiment and the needs of their current users and potential new ones.

7.3 Academic Implications

The present study made two separate academic contributions. Firstly, the study provides an addition to the current literature on freemium business models by assessing the impact of the aforementioned model on public sentiment, especially on AI platforms. Due to the fact that AI platforms is quite recent as of the date of this study, this specific aspect has been investigated in prior research studies. Moreover, the study presents the comparative advantages of LLMs over traditional machine and deep learning models for the NLP task of sentiment analysis. In addition, the study adds to the current literature that the combination of LLMs and other deep learning models did not significantly improve classification accuracy compared to LLMs on their one.

7.4 Limitations and Future Research

Although the study made some significant contribution as already mentioned, there are also some limitations. Fist of all, the dataset introduce some limitations to the study. The data

that the study relied on were obtained exclusively from one Social Media platform, namely Twitter. The dataset consisted of a sample size of 30.000 individual tweets. This amount of sample could be considered as limited with the over hype phase that ChatGPT is experiencing as of the data of this study. Additionally, it is worth noting that the representation of the overall population on Twitter may be biased due to the fact that users often selectively decide whether or not to share their opinion with a tweet based on their individual biases towards a specific topic.

Furthermore, another limitations of the study is the fact that the labelling technique employed in the dataset, may have introduced a certain degree of noise in the data, and thereby could potentially affect the quality of the data. As a result, the use of more advanced labelling methods and techniques might have produced a dataset with lower levels of noise. An additional limitation of the dataset, is the class imbalance that was observed during the analysis. This limitation as already discussed during the sentiment analysis results may be key factor why many of the models employed in the study did not have the ability to detect tweets belonging to the minority class.

Future research could potentially incorporate data from a wide range of Social Media platforms resulting in a much bigger dataset. By doing this, it could potentially enhance our understanding of user's sentiment. Furthermore, future research could potentially include data with more languages in order to gain a broader and more international perspective of the public sentiment due to the fact that the dataset used contains tweets only written in English.

Regarding the methods employed in this study, one limitations could be the simple methods used to detect the change in the sentiment over time. The methods were selected due to the time constrictions of the study, but future research could consider employing complex methods that are capable to better detect change in time series in a more detailed point. Additionally, future research could examine more combinations of LLMs with deep and machine learning models and potentially find a more efficient combination for sentiment analysis.

Furthermore, the analysis offers a comprehensive examination of the factors influencing sentiment. However, future research projects could enhance our understanding of the factors influencing public sentiment by incorporating additional data, such as demographic information. This would enable the exploration of relationships between specific age groups or education levels and topics, thereby providing deeper insights into sentiment analysis.

The final limitation of this study is that it was conducted during a period when ChatGPT

and OpenAI were experiencing an increased level of hype from the users and the world overall, which may have resulted in users overlooking the changes to the business model. It would be interesting in future researches, to examine a potential change in business model when the initial buzz has subsided.

References

[1] Y. Duan, J. S. Edwards, and Y. K. Dwivedi, "Artificial intelligence for decision making in the era of big data – evolution, challenges and research agenda," *International Journal of Information Management*, vol. 48, p. 63–71, 2019.

[2] J. C. West, "The second machine age: Work, progress, and prosperity in a time of brilliant technologies.," *Psychiatry*, vol. 78, no. 4, p. 380–383, 2015.

[3] A. M. Kaplan and M. Haenlein, "Users of the world, unite! the challenges and opportunities of social media," *Business Horizons*, vol. 53, no. 1, p. 59–68, 2010.

[4] M. Guo, "Social media competitive analysis and text mining," *Journal of Media Management and Entrepreneurship*, vol. 3, no. 1, p. 29–45, 2021.

[5] J. Hamari, N. Hanner, and J. Koivisto, ""why pay premium in freemium services?" a study on perceived value, continued use and purchase intentions in free-to-play games," *International Journal of Information Management*, vol. 51, p. 102040, 2020.

[6] T. Niemand, S. Tischer, T. Fritzsche, and S. Kraus, "The freemium effect: Why consumers perceive more value with free than with premium offers," *Proceedings of the International Conference on Information Systems 2015*, 12 2015.

[7] T. M. Wagner, A. Benlian, and T. Hess, "The advertising effect of free – do free basic versions promote premium versions within the freemium business model of music services?," *2013 46th Hawaii International Conference on System Sciences*, 2013.

[8] M. Mäntymäki, A. N. Islam, and I. Benbasat, "What drives subscribing to premium in freemium services? a consumer value-based view of differences between upgrading to and staying with premium," *Information Systems Journal*, vol. 30, no. 2, p. 295–333, 2019.

[9] S. Josimovski, L. Pulevska Ivanovska, and M. Kiselicki, "Implementing the freemium business model in the software industry: Key findings and implications," 10 2019.

[10] A. Yadav and D. K. Vishwakarma, "A deep learning architecture of ra-dlnet for visual sentiment analysis," *Multimedia Systems*, vol. 26, no. 4, p. 431–451, 2020.

[11] N. F. da Silva, E. R. Hruschka, and E. R. Hruschka, "Tweet sentiment analysis with classifier ensembles," *Decision Support Systems*, vol. 66, p. 170–179, 2014.

[12] A. Amolik, N. Jivane, M. Bhandari, and M. Venkatesan, "Twitter sentiment analysis of movie reviews using machine learning techniques.," *International Journal of Engineering and Technology*, vol. 7, pp. 2038–2044, 01 2016.

[13] C. T. Weber and S. Syed, "Interdisciplinary optimism? sentiment analysis of twitter data," *Royal Society Open Science*, vol. 6, no. 7, p. 190473, 2019.

[14] S. Tammina and S. Annareddy, "Sentiment analysis on customer reviews using convolutional neural network," *2020 International Conference on Computer Communication and Informatics (ICCCI)*, 2020.

[15] A. M. Ramadhani and H. S. Goo, "Twitter sentiment analysis using deep learning methods," *2017 7th International Annual Engineering Seminar (InAES)*, 2017.

[16] R. Man and K. Lin, "Sentiment analysis algorithm based on bert and convolutional neural network," *2021 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC)*, 2021.

[17] J. Dong, F. He, Y. Guo, and H. Zhang, "A commodity review sentiment analysis based on bert-cnn model," *2020 5th International Conference on Computer and Communication Systems (ICCCS)*, 2020.

[18] A. Bozanta, S. Angco, M. Cevik, and A. Basar, "Sentiment analysis of stocktwits using transformer models," *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2021.

[19] D. Suhartono, K. Purwandari, N. H. Jeremy, S. Philip, P. Arisaputra, and I. H. Parmonangan, "Deep neural networks and weighted word embeddings for sentiment analysis of drug product reviews," *Procedia Computer Science*, vol. 216, p. 664–671, 2023.

[20] P. Kherwa and P. Bansal, "Topic modeling: A comprehensive review," *ICST Transactions on Scalable Information Systems*, vol. 0, no. 0, p. 159623, 2018.

[21] K. Taghandiki and M. Mohammadi *Topic modeling: Exploring the processes, tools, challenges and applications*, 2023.

[22] R. Egger and J. Yu, "A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts," *Frontiers in Sociology*, vol. 7, 2022.

[23] L. Muchene and W. Safari, "Two-stage topic modelling of scientific publications: A case study of university of nairobi, kenya," *PLOS ONE*, vol. 16, no. 1, 2021.

[24] L. George and P. Sumathy *An integrated clustering and BERT framework for improved topic modeling*, 2022.

[25] J. M. Wooldridge, *Introduction to econometrics: Europe, Middle East and Africa Edition.* Cengage Learning, 2014.

[26] *Encyclopedia of Medical Decision Making.* 2009.

[27] J. Lopez Bernal, S. Cummins, and A. Gasparrini, "Interrupted time series regression for the evaluation of public health interventions: A tutorial," *International Journal of Epidemiology*, 2016.

[28] A. K. Wagner, S. B. Soumerai, F. Zhang, and D. Ross-Degnan, "Segmented regression analysis of interrupted time series studies in medication use research," *Journal of Clinical Pharmacy and Therapeutics*, vol. 27, no. 4, 2002.

[29] R. B. Penfold and F. Zhang, "Use of interrupted time series analysis in evaluating health care quality improvements," *Academic Pediatrics*, vol. 13, no. 6, 2013.

[30] G. Tunnicliffe Wilson, *Time Series Analysis: Forecasting and Control,5th Edition, by George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel and Greta M. Ljung, 2015. Published by John Wiley and Sons Inc., Hoboken, New Jersey, pp. 712. ISBN: 978-1-118-67502-1*, vol. 37. 03 2016.

[31] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, p. 273–297, 1995.

[32] R. Rifkin and A. Klautau, "In defense of one-vs-all classification," *Journal of Machine Learning Research*, vol. 5, pp. 101–141, 12 2004.

[33] C.-w. Hsu, C.-c. Chang, and C.-J. Lin, "A practical guide to support vector classification chih-wei hsu, chih-chung chang, and chih-jen lin," 11 2003.

[34] J. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Processing Letters*, vol. 9, pp. 293–300, 06 1999.

[35] H. Zhang, "The optimality of naive bayes."

[36] I. Rish, "An empirical study of the naïve bayes classifier," *IJCAI 2001 Work Empir Methods Artif Intell*, vol. 3, 01 2001.

[37] A. Mccallum and K. Nigam, "A comparison of event models for naive bayes text classification," *Work Learn Text Categ*, vol. 752, 05 2001.

[38] P. Domingos and M. Pazzani, "On the optimality of the simple bayesian classifier under zero-oneloss," *Machine Learning - ML*, vol. 29, pp. 103–130, 01 1997.

[39] H. Chen, S. Hu, R. Hua, and X. Zhao, "Improved naive bayes classification algorithm for traffic risk management," *EURASIP Journal on Advances in Signal Processing*, vol. 2021, 06 2021.

[40] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–44, 05 2015.

[41] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, pp. 2278 – 2324, 12 1998.

[42] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[43] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European Conference on Computer Vision*, 2013.

[44] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," 2014.

[45] Y. Kim, "Convolutional neural networks for sentence classification," 2014.

[46] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.

[47] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017.

[48] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," 2020.

[49] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," 2019.

[50] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," 2020.

[51] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015.

[52] M. Grootendorst, "Bertopic: Neural topic modeling with a class-based tf-idf procedure," 2022.

[53] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," 2019.

[54] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," 2020.

[55] F. Barbieri, J. Camacho-Collados, L. Espinosa Anke, and L. Neves, "TweetEval: Unified benchmark and comparative evaluation for tweet classification," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, (Online), pp. 1644–1650, Association for Computational Linguistics, Nov. 2020.

A Appendix - Examples of Tweets and Their Sentiment Classification by RoBERTa

Example Tweet	**RoBERTa Prediction**	**True Label**
maybe chatgpt just tells you what you want to hear? #lineker #chatgpt4 #chatgpt #truth_or_lie	Neutral	Neutral
maybe chatgpt just tells you what you want to hear? #lineker #chatgpt4 #chatgpt #truth_or_lie	Negative	Neutral
revolutionize your business with #chatgpt! our latest article reveals how this powerful #ai tool can help you automate customer service, generate personalized content, and optimize marketing strategies! #businessautomation #svitlasystems	Positive	Positive
"students need us to help them make sense of this powerful technology and use it creatively—not turn a blind eye to it." writes on the possibilities of chatgpt for educators and students alike #chatgpt #instructionalstrategies	Neutral	Positive
#chatgpt the publishers of thousands of scientific journals have banned or restricted contributors' use of an advanced ai-driven chatbot amid concerns that it could pepper academic literature with flawed and even fabricated research.	Negative	Negative
italy bans chatgpt over privacy concerns #news #chatgpt #ai #tech	Neutral	Negative