universität koblenz
weiter:denken
Faculty 4: Computer Science

WeST
Web Science and Technologies
Institute for Web Science
and Technologies

Assessing ChatGPT's Performance in Analyzing Students' Sentiments: A Case Study in Course Feedback

Master's Thesis

in partial fulfillment of the requirements for
the degree of Master of Science (M.Sc.)
in Informatik

submitted by
Akshay Rajkumar Sarda

First supervisor: Prof. Dr. Frank Hopfgartner
Institute for Web Science and Technologies

Second supervisor: Dr.-Ing. Stefania Zourlidou
Institute for Web Science and Technologies

Koblenz, August 2024

Statement

I hereby certify that this thesis has been composed by me and is based on my own work, that I did not use any further resources than specified – in particular no references unmentioned in the reference section – and that I did not submit this thesis to another examination before. The paper submission is identical to the submitted electronic version.

	Yes	No
I agree to have this thesis published in the library.	☒	☐
I agree to have this thesis published on the Web.	☒	☐
The thesis text is available under a Creative Commons License (CC BY-SA 4.0).	☒	☐
The source code is available under a GNU General Public License (GPLv3).	☒	☐
The collected data is available under a Creative Commons License (CC BY-SA 4.0).	☒	☐

Koblenz, 19.08.2024

. . . .

(Place, Date) (Signature)

Note

- If you would like us to contact you for the graduation ceremony,

  please provide your personal E-mail address: **akshay.sarda2k12@gmail.com** .
- If you would like us to send you an invite to join the WeST Alumni

  and Members group on LinkedIn, please provide your LinkedIn ID :
  https://www.linkedin.com/in/akshayrsarda

Zusammenfassung

Das Aufkommen großer Sprachmodelle (LLMs) wie ChatGPT hat sich auf Bereiche wie das Bildungswesen ausgewirkt und transformiert Aufgaben der natürlichen Sprachverarbeitung (NLP) wie die Stimmungsanalyse. Transformatoren bilden die Grundlage von LLMs, mit BERT, XLNet und GPT als Schlüsselbeispiele. ChatGPT, entwickelt von OpenAI, ist ein State-of-the-Art-Modell und seine Fähigkeit in natürlichsprachlichen Aufgaben macht es zu einem potentiellen Werkzeug in der Sentiment-Analyse. In dieser Arbeit werden aktuelle Methoden der Stimmungsanalyse untersucht und die Fähigkeit von ChatGPT zur Analyse von Stimmungen über drei Labels (Negativ, Neutral, Positiv) und fünf Labels (Sehr Negativ, Negativ, Neutral, Positiv, Sehr Positiv) auf einem Datensatz von studentischen Kursbewertungen untersucht. Seine Leistung wird mit fein abgestimmten State-of-the-Art-Modellen wie BERT, XLNet, bart-large-mnli und RoBERTa-large-mnli anhand quantitativer Metriken verglichen.Mit Hilfe von 7 Prompting-Techniken, die ChatGPT instruieren, wurde in dieser Arbeit auch analysiert, wie gut es komplexe sprachliche Nuancen in den gegebenen Texten anhand qualitativer Metriken versteht. BERT und XLNet übertreffen ChatGPT vor allem aufgrund ihrer bidirektionalen Natur, die es ihnen erlaubt, den gesamten Kontext eines Satzes zu verstehen, nicht nur von links nach rechts. In Verbindung mit der Feinabstimmung hilft ihnen dies, Muster und Nuancen besser zu erfassen. ChatGPT, ein universelles Modell für eine offene Domäne, verarbeitet Text unidirektional, was sein Kontextverständnis einschränken kann. Trotzdem schnitt ChatGPT in Drei-Label-Szenarien vergleichbar mit XLNet und BERT ab und übertraf die anderen. Feinabgestimmte Modelle übertrafen in Fünf-Label-Fällen. Darüber hinaus hat es eine beeindruckende Kenntnis der Sprache gezeigt. Chain-of-Thought (CoT) war die effektivste Technik für Prompting mit Schritt-für-Schritt-Anweisungen. ChatGPT zeigte vielversprechende Leistungen in Bezug auf Korrektheit, Konsistenz, Relevanz und Robustheit, außer bei der Erkennung von Ironie. Da sich das Bildungswesen mit seinen vielfältigen Lernumgebungen weiterentwickelt, wird eine effektive Feedback-Analyse immer wertvoller. Die Behebung der Einschränkungen von ChatGPT und die Nutzung seiner Stärken könnte das personalisierte Lernen durch eine bessere Sentimentanalyse verbessern.

Abstract

The emergence of large language models (LLMs) like ChatGPT has impacted fields such as education, transforming natural language processing (NLP) tasks like sentiment analysis. Transformers form the foundation of LLMs, with BERT, XLNet, and GPT as key examples. ChatGPT, developed by OpenAI, is a state-of-the-art model and its ability in natural language tasks makes it a potential tool in sentiment analysis. This thesis reviews current sentiment analysis methods and examines ChatGPT's ability to analyze sentiments across three labels (Negative, Neutral, Positive) and five labels (Very Negative, Negative, Neutral, Positive, Very Positive) on a dataset of student course reviews. Its performance is compared with fine-tuned state-of-the-art models like BERT, XLNet, bart-large-mnli, and RoBERTa-large-mnli using quantitative metrics. With the help of 7 prompting techniques which are ways to instruct ChatGPT, this work also analyzed how well it understands complex linguistic nuances in the given texts using qualitative metrics. BERT and XLNet outperform ChatGPT mainly due to their bidirectional nature, which allows them to understand the full context of a sentence, not just left to right. This, combined with fine-tuning, helps them capture patterns and nuances better. ChatGPT, as a general-purpose, open-domain model, processes text unidirectionally, which can limit its context understanding. Despite this, ChatGPT performed comparably to XLNet and BERT in three-label scenarios and outperformed others. Fine-tuned models excelled in five-label cases. Moreover, it has shown impressive knowledge of the language. Chain-of-Thought (CoT) was the most effective technique for prompting with step by step instructions. ChatGPT showed promising performance in correctness, consistency, relevance, and robustness, except for detecting Irony. As education evolves with diverse learning environments, effective feedback analysis becomes increasingly valuable. Addressing ChatGPT's limitations and leveraging its strengths could enhance personalized learning through better sentiment analysis.

Contents

List of Figures

List of Tables

List of Abbreviations

SA Sentiment Analysis

ML Machine Learning

RLHF Reinforcement learning from human feedback

AI Artificial Intelligence

CNN Convolutional Neural Network

GNNs Graph Neural Networks

GCN Graph Convolutional Networks

GAT/GAN Graph Attention Networks

ABSA Aspect Based Sentiment Analysis

AGN-TSA Attentional-graph Neural Network based Twitter Sentiment Analyzer

MSA Multimodal Sentiment Analysis

LLMs Large Language Models

GPT Generative Pre-trained Transformer

BERT Bidirectional Encoder Representations from Transformers

BART Bidirectional and Auto-Regressive Transformers

MNLI Multi-Genre Natural Language Inference

RoBERTa Robustly Optimized BERT Pretraining Approach

RP RolePlaying

CoT Chain-of-thought

LIME Local Interpretable Model-agnostic Explanation LIME

1. Introduction

1.1. Sentiment Analysis and ChatGPT

Humans are emotional beings. Sentiment Analysis (SA) represents the examination of individuals' views, attitudes, and emotional responses directed towards an entity [37]. There have been various techniques from Natural Language Processing (NLP) which have been used for sentiment analysis over the years. The evolution of SA methods can be divided in three - Lexicon based, Machine Learning (ML) and Transformer based [26]. Lexicon based methods use predefined dictionaries of words associated with sentiments. Although simple, these methods struggle to understand the context and linguistic nuances. ML based methods provided a breakthrough in SA in which models are trained on labeled datasets and are used to predict sentiments. They improved the accuracy of SA and context comprehension but require large datasets. With exponential increase in available data and various linguistic structures on the web they are not able to capture intricacies [26].

Transformer based methods, the latest of the three have brought drastic changes to the domain of SA using NLP. Their pre-training and fine-tuning approach has shown much superior performance, context awareness. In pre-training, the model is trained on large unlabeled data and in fine-tuning the models is adopted to a specific task/domain [11]. Various transformer based models are available such as BERT [11], XLNet [62], GPT [45] which use transformers but have different architectures. While BERT, XLNet are accessible via libraries, ChatGPT serves as a conversational artificial intelligence interface for GPT, harnessing NLP techniques to engage in lifelike interactions [54] that "answers follow-up questions, admits its mistakes, challenges incorrect premises, and rejects inappropriate requests "[1]. It was also introduced by OpenAI and uses reinforcement learning from human feedback (RLHF) which is a technique used to train models, in NLP, by incorporating feedback from humans for improving the model's performance [52].

In a relatively short period of time since ChatGPT was made publicly available there have been a few studies which have explored its applications in different domains including education. It can write essays [54] or even academic papers [64] even without much background knowledge. While this poses serious concerns about authenticity of the work, authors argue that this indicates the need for change in learning goals and methods. Beyond grading, Artificial Intelligence (AI) can be used to understand patterns of learning and look for areas of improvement. This is where it could be worthwhile to test the ability of ChatGPT as a sentiment analyzer. [47] conducted an in depth study of ChatGPT and suggested the application of ChatGPT to provide real time feedback to students besides other applications. There have also been some studies on ability of ChatGPT as a sentiment analyzer [59] [51] and have

[1]OpenAI Blog `https://openai.com/blog/chatgpt/`.

found ChatGPT to be reasonably good and even comparable with fine tuned models like BERT.

ChatGPT needs to be given instructions for getting responses and these are called prompts [13]. A prompt is a particular instruction or question given to a language model to direct its actions and produce the desired outcomes [18]. ChatGPT or LLMs in general are sensitive to prompts which means the output could vary considerably depending on the input [66]. There are various techniques of prompting and those have been used to test ability of ChatGPT in SA. It has shown up to 93 % accuracy on simple SA tasks like binary SC [59] [66]. However, its results on tasks requiring understanding of linguistic nuances are varied with sarcasm and irony being among the two [26]. Moreover, its worth exploring if ChatGPT can recognize contemporary language styles, for instance those involving modern abbreviations, slang words etc.

In summary, ChatGPT has been looked upon as a tool that can assist in various NLP applications including education. However, in various studies, GPT models' performance on SA in specific contexts is considered as one of the areas for future work. In this thesis, educational context is considered to further analyze the model's ability by comparing ChatGPT with some of the other state-of-the-art (SOTA) methods e.g. BERT by comparing their performances on SA on students' reviews on courses. Additionally, the studies so far have only suggested personalized learning as an area for future development. This thesis aims to address this which could be a starting point for inclusion of ChatGPT for enhanced and personalized learning. Besides, many studies so far have used binary (two classes) or ternary (three classes) for SA with ChatGPT. This thesis adds multi class scenario with five labels. Finally, with the use of various prompting techniques, this work also examines its awareness of the linguistic nuances without having to provide the entire contextual information.

1.2. Aim and Objectives

1.2.1. Aim

Evaluate the performance of ChatGPT in sentiment analysis of students' feedback on courses.

1.2.2. Objectives

1. Literature review of the state-of-the-art methods for sentiment analysis
2. Comparative analysis of ChatGPT and other state-of-the-art models for sentiment analysis and evaluation
3. Investigation using different prompts to check how well the model recognizes different subtleties in sentiments

To answer the above mentioned research objectives, this thesis work is structured as follows. In Section 2 literature review of the state-of-the-art methods in sentiment analysis as well as literature review of prompt engineering is discussed. Section 3 describes the methodology used in the work. Results are presented in section 4 followed by discussion in section 5. Finally, the work concludes in section 6. Structural outline of sentiment analysis methodologies and models explored in this thesis can be seen in Figure 1.

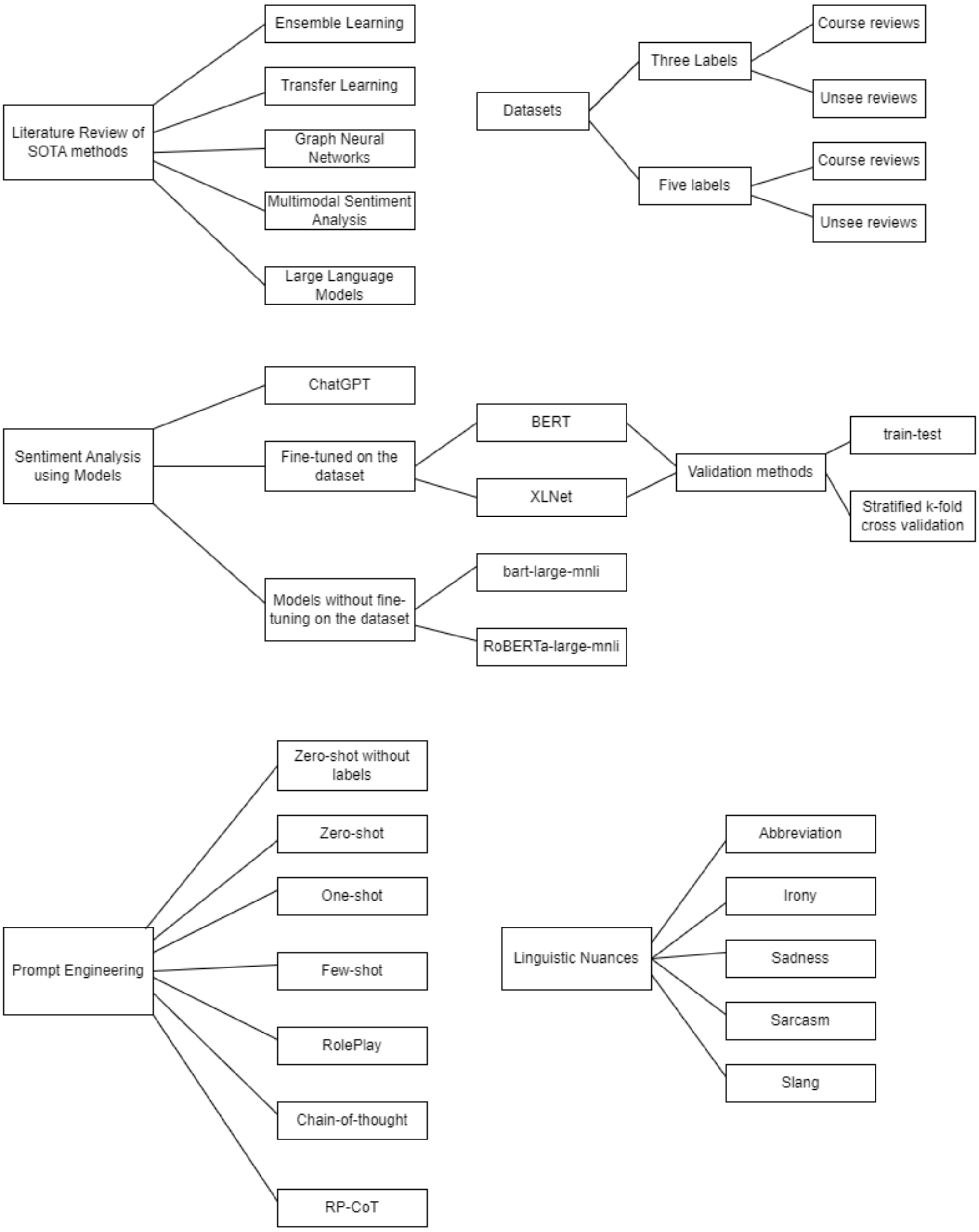


Figure 1: Structural Outline of Sentiment Analysis Methodologies and Models explored in this Thesis

2. Literature Review

2.1. State-of-the-art Methods in SA

In this section, a literature review of the SOTA methods of SA is provided. For this, recent scientific papers regarding the topic were reviewed with focus on contemporary approaches. Following five methods are discussed - Ensemble Learning, Transfer Learning, Graph Neural Network, Multimodal Sentiment Analysis, Large Language Models. Although some of these methods were introduced many years ago in machine learning, they are used in combination with newer methods such as deep learning, transformer based and hence are included in this review. They are discussed in the order in which they were introduced. For each method, its basic idea, its main components, some experimental studies using the method and their results are discussed along with advantages and limitations.

2.1.1. Ensemble Learning

Ensemble as the name suggests is a group. An ensemble of models is a collection of learning models that have some method of combining their individual predictions to produce a more broadly applicable outcome [12] [68]. The basic idea behind ensemble method is that the models in the ensemble should reduce the effect of each other's errors if the errors brought on by their biases are uncorrelated. This means that for this method to be effective it is imperative that individual models in the ensemble have some accuracy and that their biases or errors are diverse [12]. Otherwise, the errors will get compounded resulting in more inaccurate predictions. Although ensemble learning has been around for a while its use in SA is still not very prevalent. Plus, even among ensembles use of heterogeneous models is limited in current studies [25].

There are a few methods for aggregating ensembles - weighting (includes averaging and voting), meta-learning approaches, bagging, boosting, and stacking. In averaging ensemble, the class with the greatest average probability is the one that receives the final class label after the averaging ensemble computes the mean of each class's probability distributions. In majority voting whichever class gets predicted by majority of the models is chosen as the final class. In bagging (Bootstrap Aggregating) [25] , multiple instances of a base model are trained on different subsets of the training data, often created by random sampling with replacement. The predictions of these models are then combined through averaging or voting (for classification) to make the final prediction. Boosting is an iterative ensemble method where base models are trained sequentially, with each subsequent model focusing on the examples that previous models found difficult to classify correctly. The final prediction is made by combining the predictions of all the models. Stacking involves training multiple diverse models and then using a meta-model also called meta-learner to combine their predictions. The meta-learner is trained on the predictions of the base models and learns to make the final prediction based on this information [25].

Simple averaging is most commonly used in ensemble learning most of the time. But because it favors weak models, which can lower performance in many cases, the simple method is not a wise way to integrate the models. [28] proposed an enhanced approach using meta-ensemble deep learning to improve the performance of SA when using ensemble method. In their approach they used three levels of meta learners to aggregate the predictions of multiple deep model groups where meta learners are essentially responsible for learning how to effectively combine predictions. They used six sentiment benchmark datasets based on English, Arabic, and various dialects to conduct the experiments in order to assess the extended meta-ensemble deep learning approach. For every benchmark data set, sets of baseline classifiers (GRU, LSTM, and CNN) were trained, and compared the best model with the suggested meta-ensemble deep learning technique. It was observed that the enhanced approach did better than baseline models.

Many of the experiments until recently involving ensemble learning used traditional ML methods. Also, these experiments mentioned texts as word frequency based features like Term Frequency Inverse Document Frequency (TF-IDF) which don't consider the context in which the words are used [53]. [53] proposed hybrid ensemble comprising of deep learning models such as RoBERTa, LSTM, GRU. They used RoBERTa embedding which is able to record how important a word is in the provided context. In their experiment after pre-processing, textual data was passed to combinations of deep learning models to extract features and classify. Final prediction was derived by average ensemble and majority voting. The authors used three datasets viz. Internet Movie Database (IMDb)[35], Twitter US Airline Sentiment dataset[2] which came originally from Crowdflower's Data for Everyone library and Sentiment140 dataset [19]. As the data in the Twitter US Airline Sentiment dataset is imbalanced they employed data augmentation using pre-trained word embedding to ensure that all classes have same number of samples. They observed an improvement in performance of the models due to augmentation. They compared the performance of the ensemble by using different embeddings - RoBERTa, BERT and A Lite BERT (ALBERT) and found that hybrid models that used RoBERTa embeddings showed the highest accuracy above 91 %. Finally, they also compared the results of their hybrid ensemble containing deep learning models with ensembles containing just ML models and observed that the ensemble hybrid deep learning model using majority voting method did better than all other methods for all three datasets.

Ensemble learning offers increased accuracy by compensating for errors of baseline models, avoids overfitting and decreases variance and bias. These methods can improve the generalization capabilities of deep learning models, to ensure more

[2]https://www.kaggle.com/datasets/crowdflower/twitter-airline-sentiment

consistent performance across various datasets and domains [23]. High computational requirement is one of the limitations of ensemble method [23]. Combining multiple models demands robust infrastructure. Training and maintenance of distinct models at the same time can add up the computational cost and training time which could be an even bigger problem in time-sensitive demands. As the data is changing constantly, the baseline models within ensembles need to be monitored and assessed constantly which requires meticulous planning and could be resource intensive [23]. [28] identified choosing correct baseline models and deciding their number as challenges in enhanced ensemble learning. Other challenge is that as the time complexity grows along with the size of the data. Additionally, multi-class problem using this approach needs to be explored.

2.1.2. Transfer Learning

Transfer learning is a technique for taking advantage of the knowledge gained from one task or domain (Source) to enhance the performance of other task or domain (Target) [41]. [31] summarized algorithms and applications of SA using transfer learning. They discussed three types of transfer learning methods - Parameter transfer methods, Instance transfer methods and Feature representation transfer approach [41]. This method takes the advantage of pre-trained models on large dataset and shares the parameters of this model to the other model like knowledge transfer. With this, the other model can be fine-tuned for specific task like sentiment analysis. In Instance transfer method, data (instances) from source domain is reused in target domain. For instance, in cross lingual sentiment analysis [61] the data from the language for which sufficient labeled data is available such as English can be used as training data on another language which may improve the classification. This method works well when the source and target domain are similar or related. TrAdaBoost [7] is an algorithm which is one of the methods used to get instances from source domain. In feature transfer methods, it is imperative that source and destination domain have common features. For instance, if there is a large image dataset of all animals (Source) and there is a small dataset (e.g. just animals of the Cat family as Target) it could be useful to use this method. Via transformation both domains could be brought to same feature space as well [31].

Transfer learning can also be categorized as Multi-task learning and Sequential transfer learning. In Multi-task approach, source and destination tasks learning occurs simultaneously i.e. parallelly. Sequential transfer learning is based on the idea of pre-training then fine-tuning [36]. In the first step pre-training from large unlabelled data in self-supervised manner for model is performed and then it is fine tuned for a particular task based on the knowledge from the previous step [5]. This method could be useful for existing large pre-trained model e.g. BERT which is trained on large corpus but lacks knowledge specific to a domain or task.

In Visual Emotion Analysis, emotional polarity expressed in an image is identified [38]. [38] employed transfer learning in their study with an aim to enhance performance of image categorization by using VGG-19. VGG-19 [24] is a Convolutional Neural Network (CNN) which is widely used for image classification. It is trained on ImageNet [10] collection which is a dataset for images with image count in millions. [38] leveraged transfer learning and modified last layer of VGG-19 as per their requirements to classify images as positive and negative. They tested their model with three datasets - The FER2013 dataset [20], the JAFFE dataset [34], and the Cohn-Kanade Dataset (CK+) [33] which are used for studying facial expressions. They observed that their proposed model showed accuracy values of 65 %, 93 % and 99 % respectively which is improvement from other existing models. However, they also noted over-fitting as one of their finding. [4] conducted a study to see the impact of using transfer learning from a language with high resources (English) to language with low resources (Italian) in sentiment analysis. Depending on whether textual resources available of a language determine if it is a low resource or high resource language. They found that using multilingual BERT (mBERT) fine tuned on a mixed (English and Italian) dataset gave better performance than models specific to Italian language (BERT BASE Italian and BERT BASE Italian XXL).

Negative transfer learning could be a major problem in transfer learning. Here, the knowledge from source domain could have a negative effect on target domain instead of positive. Besides, all three methods discussed above have advantages and limitations. In Parameter transfer method, computational cost of training in target domain can be reduced by using parameters from pre-trained model. However, parameters from source may not easily adjust to target. In Instance transfer method, examples from source domain can enrich target and this method is easier to implement. Similarity between two domains is required for good results though. Feature representation, on the other hand, is not dependent on domain but on common features. But integrating features from both domains optimally can be challenging.[31].

2.1.3. Graph Neural Network

Graph Neural Networks (GNNs) are neural network architectures designed to process graph-structured data [65]. GNNs are influenced by CNNs. To some extent, they overcome the limitation of CNNs. CNNs are suitable for data in a grid like structure which might not be able to handle graph structures [27]. GNNs are potent tools for analyzing graph-structured data, utilizing graphs' inherent relational structure to capture intricate dependencies and patterns effectively. Two functions at the core of GNN are message passing and aggregation [1]. Message passing involves passing information between neighboring nodes in the graph where message usually contains information about the neighboring nodes' features and their relationships with the current node. After collecting messages from neighboring nodes, each node aggregates these messages to form a summary of the information from its local neighborhood. Aggregation helps in giving an idea about the neighbourhood of a node and its relationship with the nodes in the neighbourhood. [27] described three types of GNN models viz. Graph Convolutional Networks (GCN), Graph Attention Networks (GAT/GAN), and GraphSAGE models. Out of these GCN is discussed above.

A text could contain different expressions of different polarity (positive, negative or neutral). In Aspect Based Sentiment Analysis (ABSA), sentiment analysis is performed to assign polarities of each aspect [63]. [57] reviewed studies related to Twitter Sentiment Analysis(TSA) where primary task is to analyze tweets of the users to reach a conclusion. They observed that all those studies relied solely on tweet-texts without taking users' network (their connections) into account. As graph is a suitable method to represent social network like Twitter, the authors also reviewed studies on GNN. They proposed Attentional-graph Neural Network based Twitter Sentiment Analyzer (AGN-TSA) in order to leverage architecture of GNN and combine tweet-texts and user connection to perform TSA. It has three main layers First, the Word-Embedding Layer, after preprocessing the text this layer gives numerical representation (vector) that captures its meaning. Second, the User-Embedding Layer converts the tweet-text data representation (from first layer) into user embeddings. For this, for each user a word frequency sequence is generated which shows frequency of each word used by the user. The final layer, the Attentional-Graph Layer does coupling of the user-embedding and user's connection information to create a new user representation.

They tested their proposed framework for analyzing the tweets related to US presidential election of the year 2016 and data was obtained using APIs offered by twitter. For this they chose 1224 users who had expressed their views about the topic at least once during the period when the data was being collected and had at least one interaction on twitter with the other users in the dataset. Thus, these fulfilled the requirement of AGN-TSA. They compared the results of AGN-TSA with traditional methods of SA using the same dataset and observed 5 percent increase in accuracy. [8] studied one of the potential areas of applications of SA viz. stock market. GNNs represent financial data graphically, with stocks as nodes and company relationships as edges, employing convolutional methods to capture spatial correlations within the graph. GNN models with recurrent structures tailored for changing times enable modeling of evolving stock relationships over different time intervals, bolstering their capacity to capture dynamic market trends. SA enhances stock prediction by unveiling market sentiment, investor behavior, and emotional impact on prices. Integrating it into models offers a competitive edge, facilitating better comprehension of market trends and informed investment decisions [8]. They explored the potential role of GNN in SA and highlighted the positives of it.

Ability to capture complex dependencies is the core feature that GNNs offer. They can produce more accurate results compared to traditional methods of SA. More specifically, in the dynamic domain of stock market they provide a promising approach to update predictions constantly with changing factors. However, a big challenge for GNN is that GNNs heavily depend on the structure of the training graph, which can pose challenges when adapting as the graph structure and data changes [1]. Moreover, in case of stock market sentiment analysis to predict something, if the data contains noise, incorrect information or prejudices it can lead to incorrect analysis and hence predictions. Processing resources required for GNNs are quite high which limits their use if such resources are not available.

2.1.4. Multimodal Sentiment Analysis

Traditionally SA has always been based on texts. An opinion about an object is also available in the form of videos, audios, images etc. Multimodal Sentiment Analysis (MSA) takes these other modalities of communication into consideration for SA. It could be bimodal (involving two of these modalities) or trimodal (three of these modalities). Fusion is at the core of MSA. It involves merging, filtering, and retrieving the necessary features from the data gathered via a variety of modalities [15]. There are two main parts in the process of MSA. First is extraction of features from each modality. Second, passing these features to a fusion model which predicts the sentiment from these features. There are various techniques/libraries which are used for extracting features from each modality. For textual features word embedding technology is used which provide vector representations of the words. Word2Vec [39], Glove [42] are the most commonly used methods used for this. BERT which is transformer based model is also used due to its capabilities of parallel processing while retaining contextual information.

Visual feature extraction refers to extracting features from videos/images using facial expressions or body postures. Public libraries like Computer Expression Recognition Toolbox (CERT)[30] are used for this. For extracting features from audio which is also known as acoustic feature extraction open source softwares or libraries are most commonly used. e.g. OpenEAR[14] calculates acoustic features like prosody, energy, vocal probability on its own [69]. Most commonly used method for SA using audio is to convert audio/speech to texts and perform SA on it whereas analyzing sentiment directly from the spoken words is a relatively new area [9]. Early fusion, also known as feature-level fusion, combines data from all modalities (text, audio, visual) right at the input stage by merging feature vectors into one input for a classifier like an SVM or neural network. On the other hand, late fusion, which is also known as decision-level fusion, processes each modality independently to make sentiment predictions and then combines these predictions using methods like averaging or voting.

Because of multiple modalities MSA can capture emotions with more granularity.

However, it comes at the cost of additional complexity and computational cost. Early fusion Simplifies the model design which means it needs to focus only on the classifier. However, merging data early can miss out on the detailed and nuanced information from each modality, which can lead to overfitting. As the merging happens at the decision level, In late fusion details of each modality can be handled and it is adaptable to changes in modalities. But low-level interactions between different types of data are not there, missing some nuanced cross-modal insights [17] [69] .[15]. Intra-modality dynamics and inter-modality dynamics are two of the challenges in MSA. These are particularly more challenging when SA is to be done on spoken language. For instance, a sentence could have completely different sentiment depending on the facial expression a person has while saying it. Similarly, using words which are mainly used in colloquial manner (e.g. hmm, yeah followed by a pause, okay.. etc) present a challenge for multimodal sentiment analysis [9].

[15] reviewed latest approach used in MSA and various applications domains of it. They also noted the results of different studies where MSA was performed by using traditional machine learning methods and deep learning methods on different dataset and deep learning based methods showed comparable and in some cases better performance.

2.1.5. Large Language Models

There have been methods to process sequential data like texts which have been around for a while e.g. Recurrent Neural Networks(RNNs) [22] and Long-Short Term Memory (LSTM) [21]. However, for longer texts they not only begin to loose context but also are slower due to their sequential nature [2]. Large Language Models (LLMs) are the models that can overcome these problems. Language models are computational models that can comprehend and generate human language. LLMs are advanced LMs and are called LLMs due to their massive input parameters size[6]. The main component that LLMs are based on are Transformers [6], which have have brought about a paradigm shift in SA [26]. Transformer is an architecture in deep learning which was proposed by Google and built on the concept of Attention [55]. Besides this, two major components of transformer are Encoder which encodes the input into a series of representations, Decoder, which generates the output sequence [55]. Attention is based on the idea that each word in a text has a different importance relative to other words in given context called positional encoding [55] [2]. Transformers have enabled LLMs to revolutionise the NLP tasks like SA. GPT, BERT and XLNet are examples of SOTA and some of the most and popular LLMs.

GPT. GPT which stands for Generative Pre-trained Transformer is a model introduced by OpenAI. It is the auotoregresisve model which means it has the ability to predict the next word in the sequence. It is a generative model and can generate coherent, contextually relevant texts and has been trained on massive amount of data. There have been multiple versions of GPT which have been released with each having different parameter size out of which GPT-3 has 175 billion parameters [26].

BERT. BERT stands for Bidirectional Encoder Representations from Transformers and was introduced by Google [11]. The main difference between GPT and BERT is that BERT is bidirectional which means it reads the text from both left and right to understand the context. Two core concepts in BERT are Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) where MLM randomly masks a word in a text to predict in the given context and NSP predicts if a sentence follows another [11]. Two main variants of BERT are available viz. BERT-base and BERT-large with 110M and 340M parameters respectively [43].

XLNet. XLNet was proposed by [62] and is a Generalized Autoregressive Pretraining for Language Understanding method. Although XLNet works in autoregressive and bidirectional manner the main difference from BERT is it uses permutations by rearranging the order of words in a sentence multiple times and learning to predict each word based on all possible orderings, which helps it understand context from both directions. In contrast, BERT's MLM learns bidirectional context only from the masked positions [43]. The term XL in the word XLNet is due to the Transformer-XL model on which this model is based.

bart-large-mnli. [29] proposed BART which stands for Bidirectional and Auto-Regressive Transformers. It is built for sequence-to-sequence tasks and combines the strengths of both BERT and GPT with the use of bidirectional encoder and left-to-right decoder respectively. bart-large-mnli[3] is a variant of BART which has been fine-tuned on Multi-Genre Natural Language Inference (MNLI)[4] dataset. This dataset is a large corpus of crowdsourced texts of different genres and widely used for training models. bart-large-mnli which is large as it is trained on 407M parameters is suitable for zero-shot(without fine-tuning) SA of various classes because of the MNLI.

RoBERTa-large-mnli. RoBERTa stands for Robustly Optimized BERT Pre-training Approach and was proposed by [32]. It is robustly optimized version of BERT as it is trained on much larger dataset, uses dynamic masking instead of static with better performance. RoBERTa large is the large version of RoBERTa with 356M training parameters. RoBERTa-large-mnli[5] is RoBERTa-large fine-tuned on MNLI. Like bart-large-mnli, this is what makes it suitable for zero-shot SA.

Hyperparamter tuning. A model is trained using learning algortihms and when it

[3]https://huggingface.co/facebook/bart-large-mnli
[4]https://huggingface.co/datasets/nyu-mll/multi$_{n}li$
[5]https://huggingface.co/FacebookAI/roberta-large-mnli

is trained for a specific task there are certain configurations which govern the training process. These are called hyperparameters which learning algorithm learns at the end of the process and they affect the model performance[6]. Before the training phase, aim is to identify a set of hyperparameter values that achieve the best performance on the data within a reasonable timeframe. This process is known as hyperparameter optimization or tuning [60]. There are two approaches for selection of best hyperparameters - manual and automatic. In manual method, different combinations of hyperparamters are tried which is very inefficient in case of high number of hyperparameters. In automatic methods there are mainly three methods - Grid Search, Cartesian hyperparameter search and Bayesian Optimization. Out of these Bayesian Optimization is considered to be better than the others. It selects the most promising hyperparameters to test and learns from previous results, saving time and computational resources [62].

There have been various studies exploring the performance of transformer based models in the last few years on popular datasets. [26] compared variants of GPT models with other high performance models which were used previously. They used Semeval-2017 task 4 [49] and found the accuracy of some variants of GPT to be more than 95 %. [43] compared pre-trained models like BERT, its variations, XLNet and T5 on binary sentiment analysis and observed accuracy higher than 90 % on 5 out of the 6 models with value up to 96 % in XLNet.

There are significant advantages that LLMs offer over traditional methods. Firstly, as opposed to sequential processing, they can process each part of the text at the same time and this parallellization improves the speed massively. They can handle long range dependencies which means that even in long texts they don't loose the context and understand the meaning of a word by considering all its surrounding context at once [43] [6]. Moreover, their applications in LLMs have exhibited their scalability and versatility in applications which includes not only SA but also complex NLP tasks. Along with their benefits, transformers also have certain challenges. Firstly, they are computationally expensive. e.g. fine tuning or pre-training LLMs like BERT, GPT requires GPUs/TPUs with high processing power and can take a long processing time. For these models to work effectively they require very large datasets during training which may not be available in the given domain or language. Additionally, interpreting the results of the transformer based models is challenging. Despite their impressive performance their functioning has black box nature. Understanding how a model made a decision is difficult which could be a problem especially in applications where the explanation is critical [26] [2].

The methods discussed in this section are summarized in the below Table 1.

[6]https://towardsdatascience.com/parameters-and-hyperparameters-aa609601a9ac

Method	Description	Key Components	Strengths	Weaknesses	Accuracy on Experiments
Ensemble Learning	Combines multiple models to improve performance	Multiple base models, combination strategy (e.g., voting, stacking)	High accuracy, robustness	High computational cost, complexity in implementation	High
Transfer Learning	Utilizes pre-trained models and fine-tunes them on specific tasks	Pre-trained models and fine-tuning	Saves training time and computational cost, low resource languages can benefit	Negative transfer learning, May not generalize well if source and target tasks differ	Varied
Graph Neural Networks (GNNs)	Leverages graph structures to capture relationships	Nodes, message passing, aggregation	Capture complex dependencies	High computation cost, dependence on graph structure	High
Multimodal Sentiment Analysis	Combines different modalities (e.g. text, image, video) for better analysis	Fusion methods and multimodal data	Can capture emotions with more granularity	Resource intensive, Intra-modal and inter-modal dynamics	No primary experiment results, secondary results are high
Large Language Models (LLMs)	Language models trained on massive datasets	Transformers	Speed, high contextual understanding, versatility for applications	Need large datasets, very resource intensive, potential biases	High to Very high

Table 1: Comparison of State-of-the-Art Sentiment Analysis Methods

2.2. Prompt Engineering

Literature review of Prompt Engineering is discussed in this subsection. Starting with the basic information of prompts, types of prompting/prompting techniques, some experimental studies using different prompting techniques along with challenges in effective prompting and recommended practices are discussed. With the advent of LLMs a new discipline called Prompt Engineering has emerged [3]. Prompt engineering is a process in NLP that involves designing, refining, and optimizing prompts to effectively communicate user's purpose to language models like ChatGPT. The goal is to enhance the model's performance on specific tasks by carefully crafting the input prompts [13] [56].

2.2.1. Prompt Techniques

[56] categorized prompts into two viz. manual and automated. Manual prompts are meticulously designed by humans whereas automated ones are created by using algorithms and automated methods. Manual prompts can be Zero-shot or Few-Shot while automated prompts can be categorized as Discrete and Continuous. In Zero-

shot prompting, the model is manually provided with instruction in natural language without a lot of details. It relies on the knowledge of the pre-trained model and its ability to learn from the context it identifies while responding. In few-shot prompting on the other hand, a small number of examples are provided as guidance for the model to achieve better performance. Discrete prompts involve automatically finding fixed templates or specific phrases in natural language to help guide the model in generating the desired response, whereas, instead of using fixed templates or questions, continuous prompts operate in a more abstract space called the "embedding space." This means they don't have to be in readable text form. They can adjust and fine-tune certain settings based on training data, which allows them to be more adaptable and efficient. Both manual and automated methods have their pros and cons. Manual prompts give more control with possibility to provide precise details. However, they can be time consuming to design. Automated, can be adaptable and efficient but their effectiveness depends on the search algorithms. Besides, the above mentioned manual prompts there are also prompts called One-shot prompts [7] which is a technique for generating responses using input that includes minimal additional information. As the name suggests, 'one-shot' refers to providing just a single example or a single template as the supplementary information.

[58] studied the enhancement of LLMs' performance in SA through the application of prompting strategies. They discussed two prompting strategies - RolePlaying (RP) and Chain-of-thought (CoT). RolePlaying involves guiding a language model by assigning it a specific role or persona. e.g. "You are an expert of Python programming". Chain-of-thought prompting is a recent technique that encourages large language models to explain their reasoning with steps in decision making. The method involves either explicitly instructing the model to provide reasoning or providing it with a few examples that include detailed reasoning by following which, the model learns to show its own reasoning process when answering similar prompts.

2.2.2. Related Work

[1] studied the responses of ChatGPT based on prompt sensitivity which means how the response changes as the prompts change. They performed their experiments on the affective computing tasks like SA, toxicity detection, and sarcasm detection. Multiple types of prompts were used like Zero-shot (called Base in their study), RolePlaying (Expert), Ignorant (confusing prompt), CoT and various combinations of all these.Two of the key parameters in decision making of ChatGPT viz. Temperature and Top-p were also included. Temparature means the randomness of output with values between 0 and 1 (1 means extremely random i.e. almost always a different response even for the same prompt and 0 means almost no randomness). ChatGPT selects the next word in its output from its probability of occurring. The top-p controls the probability value that it should consider as a threshold. It was found

[7] https://promptsninja.com/few-one-zero-prompting/

that the model is sensitive to the two parameters mentioned and CoT prompts produced best results. Moreover, irrelevant or confusing input worsens the output. [58] conducted SA using ChatGPT (GPT-3.5) on three datasets - IMDB dataset[8], FiQA (Financial Phrase Bank datasets)[9] and Amazon Reviews[10]. They used four strategies - Vanilla prompting (straight instruction), RP, CoT and RP-CoT and found RP-CoT produced the best results in terms of accuracy with the values of 92 %, 83 % and 94 % on the datasets from the domains of movie, finance and shopping respectively. The SA task becomes more challenging for the model as the number of labels (classes) in the dataset increase.

Use of any specific prompting technique does not guarantee best results. [26] performed SA on benchmark dataset SentitEval 2017. They carefully designed zero-shot as well as RolePlaying and Zero-shot combined prompts. They compared the results with other high performing models (e.g. RoBERTa) and found GPT-3.5 Turbo (a variant of GPT) to be performing better by 25 % from 72 % to 97 %. [66] examined capabilities of ChaGPT in SA at various levels like Document level, Sentence level, and tasks like Sentiment Classification (SC), Aspect Based Sentiment Analysis (ABSA). SC refers to task of classifying a text e.g. as Positive or Negative. ABSA, on the other hand is more fine grained and targets a specific aspect within the text. They found LLMs achieve very good results on simple SA tasks even on zero-shot prompting. While it showed varying accuracy on different datasets in SC, with as low as 48 % on SST-5 [50] and went as high as 97 % on Yelp-2 [67] dataset. They tried few-shot prompting by providing increasing number of examples and compared their performance with Small Language Models (SLMs) on the same prompts. SLMs are trained on small number of parameters usually in a specific domain. For simple SC task, number of shots don't impact the result significantly. However, for ABSA, the result improves. Moreover, if the text is too long the performance does not get better even with higher shots. This could be due to the limitation of LLM to deal with context in a text that is extremely long.

[18] focused on the importance of prompt engineering in academic writing in their study and pointed out some challenges in prompting some of which are applicable to prompting in any domain and not just academic papers. Ambiguity refers to a prompt that lacks specificity e.g. The word Green could have more than one meaning (e.g. color, environment friendly). This is similar to another challenge called lack of context. Each LLM may have a bias depending on the data that it was trained on. Additionally, if the user prompt insinuates a bias e.g. gender bias the response would also be influenced by that. Designing prompt but having unrealistic expectations from the model or assuming that it knows something could lead to bad responses. [56] provided a comprehensive and systematic overview of prompt en-

[8]http://ai.stanford.edu/ amaas/data/sentiment/

[9]https://www.kaggle.com/datasets/sbhatti/financial-sentiment-analysis

[10]https://www.kaggle.com/datasets/bittlingmayer/amazonreviews

gineering methods for NLP tasks in the medical domain and listed a few challenges. The major challenge in medical domain which could be present in other domains as well is Data scarcity. Plus, medical field involves use of specific terminology which the model may not be aware of. Finally ethical consideration is a challenge which is a commonly mentioned in the above studies. It's up-to the user to design prompts devoid of unethical or illegal intent.

[13] suggested best practices for better responses and user experience. Firstly, iteratively modifying the prompts based on responses helps to fine tune model's behavior. ChatGPT can be very creative. It is vital to adjust the creativity at the right level to match the intended response. There could be scenarios where real time data is required for ChatGPT to perform some action e.g. weather report or stock prices. In this case, an external service like API could be used to fetch the real time data and provide it to ChatGPT to take some action like analysis. [16] introduced the the term Promptgramming which means programming prompts for generative AI models like ChatGPT. He provided key principles of effective promptgramming viz. Specificity (prompts should be clear instead of vague), Contextualization (provide contextual information for model to understand the prompt and respond better), Step-by-step instructions (break the complex instructions into steps), Iterative refinement (adjust the prompts based on response iteratively). Writing effective prompts might be believed to be an intuitive skill. [40] conducted a study in the domain of art. They invited participants to judge the quality of the prompts, write textual prompts to generate images and improve their prompts after getting the output from initial writing. It was concluded that effective prompt writing is a skill that must be acquired and requires knowledge of keywords and key phrases.

2.2.3. ChatGPT in Education

In a relatively short period of time since ChatGPT was made publicly available there have been a few studies which have explored its applications in different domains including education. [54] conducted a qualitative instrumental case study to investigate the utilization of ChatGPT within the realm of education, specifically among early adopters. It was a three stage study one of which involved analysis of the views of the people who used ChatGPT in their education. One of the revelations from this exercise was the need for innovative teaching and learning approaches. For instance, with ChatGPT at disposal essay composition should not pose a formidable challenge to students, including those devoid of prior knowledge in a given subject matter. The authors emphasize the need for future research to strike a balance between using chatbots and ensuring meaningful human interaction and feedback in education, which can ultimately yield superior results for both educators and learners.

In a separate study by [64] it was hypothesized that ChatGPT has the potential to bring about shifts in educational learning objectives, the nature of learning tasks, and the methods of assessment and evaluation. In the study, the author asked ChatGPT to write an academic paper, titled Artificial Intelligence for Education and found that it was structured, partially accurate, fast and required limited professional knowledge of the topic by the author. While this suggests a danger that students may outsource writing work to AI, the author argues that it is important to set the learning goals and methods appropriately such that they revolve around creativity and critical thinking of the students. Beyond grading, AI can be used for understanding trends and patterns of learning and looking for areas of improvement. This is where it could be worthwhile to test the ability of ChatGPT as a sentiment analyzer.

2.2.4. Research Gap

Although ChatGPT has been looked upon as a tool that can assist in education, the studies so far have only suggested personalized learning as an area for future development. Because of its ability of NLP, sentiment analysis is viewed as ChatGPT's useful application. However, its application for sentiment analysis in education domain is largely unexplored which could be beneficial in personalized learning. This thesis aims to address this research gap which could be a starting point for inclusion of ChatGPT for enhanced and personalized learning.

2.3. Research Questions

With the aim to evaluate the performance of ChatGPT in sentiment analysis of students' feedback on courses this thesis work seeks answers to following research questions.

1. What are state-of-the-art methods for sentiment analysis?

   Literature Review on Sentiment Analysis Methods: Thorough literature review was conducted to identify and analyze state-of-the-art methods in sentiment analysis. This involved examining the latest advancements, techniques, and theoretical approaches in the field, providing a comprehensive overview of current methods/models.

2. How accurately does ChatGPT perform sentiment analysis on student feedback regarding courses compared to other state-of-the-art models, such as XLNet and BERT?

   Comparative Analysis of Sentiment Analysis Models: Perform a comparative analysis of ChatGPT with other state-of-the-art models in sentiment analysis, such as XLNet and BERT, bart-large-mnli, RoBERTa-large-mnli. This includes evaluating their performance, accuracy, and applicability in different sentiment analysis contexts, with a focus on understanding their strengths and limitations.

3. How well does ChatGPT understand nuances in the texts when applied in sentiment analysis?

   Prompt Engineering and Nuance Recognition in Sentiment Analysis: Engage in prompt engineering to create a diverse set of test prompts, aimed at evaluating how effectively ChatGPT recognizes and interprets different subtleties in sentiments. This research question seeks to explore the model's proficiency in discerning complex emotional nuances, such as sarcasm, irony, and subtle emotional tones, through carefully crafted prompts that challenge its interpretative capabilities in various textual scenarios.

3. Methodology

The main objectives of this thesis are to evaluate ChatGPT on SA by comparing with other SOTA models and assess its ability to understand linguistic nuances. This section describes the research design and methods used to achieve these goals. The details of the research design, datasets, pre-processing, SOTA models chosen, fine tuning of the models, prompting techniques used and evaluation metrics are given here. This will provide a comprehensive framework for answering the research questions.

3.1. Research Design

In order to provide comprehensive overview of the capability of the models used, the methodology for this thesis was mixed that will include both quantitative and qualitative methods.

Quantitative. This involves calculation of numerical data i.e., metrics to evaluate the performance of a model such as Accuracy, Precision, Recall, F1-Score etc. to assess how well do the models categorize the text as per the sentiments. Additionally,this helps for comparative analysis of ChatGPT, XLNet and BERT, RoBERTa, BART using the same metrics. The comparison makes it easier to understand how ChatGPT performs compared to other approaches.

Qualitative. This covers assessing ChatGPT on those factors which are non-numeric. This includes examining how well does the model understand and interpret the context and different types of emotions and subtleties (e.g. sarcasm, irony etc). This lead to also indicate where the model fails or struggles which helped to determine its limitations.

This mixed approach was suitable for this work as the quantitative methods facilitate to contextualize the performance of ChatGPT and other models. Qualitative methods allowed to gain deeper understanding of its knowledge of linguistic nuances.

3.2. Data Collection

There are two types of datasets used in this work. One for quantitative and the other for qualitative analysis. For quantitative analysis, a kaggle dataset that is called '100K Coursera's Course Reviews Dataset'[11] that contains reviews of students on courses available on an online platform called Coursera[12] was used. It contains 100k textual reviews of various courses scraped from the website by the author. There are

[11] https://www.kaggle.com/datasets/septa97/100k-courseras-course-reviews-dataset
[12] https://www.coursera.org/

two files in the dataset. One in which reviews are grouped by courses and the other in which they are not. For this work, the latter (reviews.csv) is used, as the grouping is not essential. It contains three columns viz. Id, Review and Label where Id is the serial number, Review is the text of the review and Label is the actual label.

This dataset was chosen as its educational domain is relevant for this thesis work. There are five labels possible for each review - 1 (Very Negative), 2 (Negative), 3 (Neutral), 4 (Positive), 5 (Very Positive). The number of available datasets in this domain which could work as standard datasets was researched. The dataset with 5 labels were even fewer. The license for this dataset as mentioned on the platform was Open Data Commons Open Database License which means it was available for free use. It is a popular dataset with a few thousand downloads. Its usability of 7.06 out of 10 is slightly on the lower side mainly as some of the metadata about it is not provided by the authors. Finally, although 10 % of the total reviews were considered in this study due to hardware constraints of the fine tuning process of models, the sheer volume of reviews covers diverse types of texts with several thousand words and styles of the modern users on the web.

For qualitative analysis data was collected from different sources for each linguistic nuance that this thesis explores viz. Irony, Sarcasm, Sadness, Abbreviations, Slang. This is because there was no single dataset found which contains labelled texts for all the nuances. For the text containing Irony and Sarcasm, a kaggle dataset called Tweets with Sarcasm and Irony[13] containing labelled texts was used. It has two csv files and the texts used in this work are from the train file. For texts containing sadness a different kaggle dataset called Emotion Dataset for Emotion Recognition Tasks[14] was used. This dataset has three files and the texts chosen for this work are from the csv file named validation. Both these datasets have texts which are messages on twitter (now called X), have a high usability (8.24 and 10.00 respectively), and were open for use via the CC0 license. For the remaining two linguistic nuances, no standard datasets with labels assigned were found. For this reason, the texts were taken from language learning platforms on the web. For slang, the examples from platform called DoTEFL[15] were taken. These are the slang phrases mainly used in American English. For abbreviation, the examples were taken from SimpleTexting[16] and Busuu[17]. All the examples were publicly available at the time of writing this work. 50 examples of each nuance were chosen from these datasets. These include variety of words with all nuances suitable for testing.

[13]https://www.kaggle.com/datasets/nikhiljohnk/tweets-with-sarcasm-and-irony
[14]https://www.kaggle.com/datasets/parulpandey/emotion-dataset
[15]https://www.dotefl.com/american-slang-words/
[16]https://simpletexting.com/blog/text-abbreviations/
[17]https://www.busuu.com/en/english/abbreviations

3.3. Data Analysis Methods

3.3.1. Technical Resources

Software and Tools. Python is the primary programming language used in this thesis work for various tasks like pre-processing, Exploratory Data Analysis, fine tuning of the models and visualizations of the results. Kaggle Notebook which is an interactive, cloud-based environment provided by Kaggle was used for coding. Fine tuning large language models is a very resource intensive task. GPU P100 with CUDA and large memory capacity is provided by Kaggle as accelerator which was utilized in this work. ChatGPT was accessed via the website[18]. A paid subscription of the ChatGPT with access to latest models was used in this work.

Libraries. As Python was the programming language used in this work, various python libraries were used. Transformers library, provides functions/classes for fine-tuning and training models. PyTorch is a deep learning framework used for building and training neural networks. sklearn.model_selection was used for train-test split and stratified K-fold cross validation. Optuna, LIMETextExplainer were used for hyperparameter optimization, interpretation of predictions in case of misclassification respectively. Besides these, pandas, numpy, re, matplotlib were used for various tasks of data handing and modeling.

3.3.2. Exploratory Data Analysis

This step was carried out to get a better understanding of the underlying structures, pattern and characteristics of the data.

Label Distribution. This shows the distribution of labels (sentiments) in the dataset. It helps to see how balanced or imbalanced the dataset is.

Text Length. This shows the frequency of the texts of different lengths. This can help to explore how models perform on smaller or longer texts.

Word Cloud. This displays words with their size corresponding to their frequency or importance in the dataset.

3.3.3. Pre-processing

The transformer based large language models like BERT, XLNet, GPT are equipped to understand diverse texts without the necessity of a lot of pre-processing. Nevertheless, some of the cleaning required for this work are mentioned below.

[18] https://chatgpt.com/

English Texts. It was observed that the dataset for SA contains a very small percentage of texts which are not in English language. As this study focuses on English language texts those texts were removed. For this, Google sheets which provide feature to detect the language used in a text was used.

Duplicates Removal and Subset Selection. All the duplicates from the dataset were removed in this step to avoid redundant text due to duplication. Fine tuning of large models like BERT, XLNet is computationally very expensive which could take several hours. Considering the available time and resources a subset of the entire dataset after previous step was selected for further work with (almost) same proportion of each class in the subset.

Unintelligible Characters Removal. Unintelligible characters which are irrelevant for the analysis if present were removed from the texts of SA dataset using regular expression which included allowed characters. For prompt engineering task, the texts for sarcasm and irony dataset contained hashtags indicating the nuance which was used in the text e.g. Irony. They were removed as it would negate the purpose of testing ChatGPT's ability.

Three and Five Labels. The dataset for SA has five labels - 1 (Very Negative), 2 (Negative), 3 (Neutral), 4 (Positive) and 5 (Very Positive). In this thesis work, another copy of this was created and the number of labels was reduced from 5 to 3. This was done to compare the results in case of both 3 labels and 5 labels scenarios. The labels Very Positive and Very Negative were replaced by Positive and Negative respectively for the texts that had those labels.

3.4. Sentiment Analysis

3.4.1. Model Selection

There are five models which are selected for comparison of SA in this work viz. GPT-4, BERT, XLNet, bart-large-mnli and RoBERTa-large-mnli. As this work mainly aims to compare ChatGPT's performance, the selection of GPT was essential. GPT 4 with significantly higher number of parameters and improved ability to understand texts was chosen. BERT and XLNet were chosen as all three are comparable with few similarities and differences. All three are trained on massive amount of data and based on transformer. Yet, their data sources and architectures are distinct. Moreover, both BERT and XLNet have shown superior performance in SA by studies. However, both these models need to be fine tuned on our dataset for comparison and there are no fine-tuned versions of base models readily available with 5 labels in SA at the time of this thesis work. For this reason, two more models which are based on these LLMs and fine tuned on different datasets but suitable for downstream task of SA using 3 as well as 5 labels without needing fine-tuning for comparable performance were selected. All these models are briefly discussed in the literature review of this work. The Table 2 below shows the summary of model selection.

Base	Model	Rational
GPT	GPT-4	GPT is essential for the study and GPT-4 is SOTA with highest parameters in GPT family
BERT	bert-base-cased	BERT is SOTA and comparable in terms of architecture and suitable after fine-tuning
XLNet	xlnet-base-cased	XLNet is SOTA and comparable in terms of architecture and suitable after fine-tuning
BART	bart-large-mnli	The model is SOTA with suitable for comparison without fine-tuning
RoBERTa	roberta-large-mnli	The model is SOTA and suitable for comparison without fine-tuning

Table 2: Model Selection

3.4.2. Models Validation

Model validation is a process to evaluate the performance of a model. It can be done for various purposes. It helps to know how a model performs when tested with data that it has not seen, examine the generalization ability, improve the predictive performance [46]. There are various techniques for model evaluation. Two of these are used in this thesis which are briefly described below.

Train-Test method. In this method the dataset is simply split into train and test and is also referred to as hold out. The model is trained on the train part of the dataset and tested on the test part of the dataset. Simplicity is a major advantage of this method [46]. Although this method is not the best method for datasets which are imbalanced, this method was chosen in this work for comparison with results of other method. The split of 80-20 was set which means 80 % of the data was used for training and remaining 20 % for testing. train_test_split function provided by sklearn.model_selection model in Scikit-learn[19] library was used.

Stratified K-fold Cross Validation. This is a variation of k-fold cross validation technique and abbreviated as SKF where cross validation broadly refers to cross over of train and test. This means both train and test sets are part of both splits. K in k-fold refers to number of splits. Stratified means the distribution of sentiments in each fold remains the same as original dataset. This method is suitable for datasets which are imbalanced [46]. In this work StratifiedKFold provided by sklearn.model_selection model in Scikit-learn library was used. Considering the time and resources constraints number of fold chosen were 2. Both of the functions from Scikit-learn have a parameter called random_state and its value was set to 42 for reproducibility of the results.

[19] https://scikit-learn.org/stable/

Useen Data. After completing the fine-tuning with stratified K-fold cross validation the models were tested with the unseen data. This data was taken from the original dataset of 100k texts as only 10 percent of the data was used in the actual fine-tuning.

3.4.3. Hyperparameters Tuning

Following hyperparameters were used in this work while fine-tuning the models BERT and XLNet - Learning rate, Batch Size, Epochs[20], Weight Decay[21].

Learning Rate. It determines how much the model's weights are adjusted with each step during training. In simple words it is the speed at which the model learns. A low value will slow the training process and high value may not help the model optimize. It was set between 1e-5, 1e-4.

Batch Size. The number of training samples processed at a time is called a batch. Per Device Train Batch Size refers to the batch size per device (GPU) during training. Due to limitations of the hardware resources the value were chosen from 8,16 and 32 (only for BERT).

Epoch. Epoch indicates how many times a the model passes through the data. Too many epochs can result in overfitting. Range was set for this for each evaluation method. For train-test method the range was 2-5 and for Stratified K-fold cross validation it was 2-3.

Weight Decay. It is a form of regularization that helps prevent the model from overfitting. It was set between 1e-6 and 1e-2.

Optuna. Optuna[22] is an open-source hyperparameter optimization framework to automate the search for the best hyperparameters. It integrates seamlessly with PyTorch and expects number of trials from the user. After providing initial values of the hyperparameters as mentioned above, Optuna found the best hyperparameters after completing number of trials. For train-test method the trials were set to 5 and for the other method the number was set to 3.

3.4.4. Evaluation Metrics

Following evaluation metrics[23] (e.g. accuracy, precision, recall,F1-score) were calculated and a confusion matrix was derived using the programming language Python.

[20] https://medium.com/@poojaviveksingh/hyperparameter-tuning-in-machine-learning-a39c8fafe6ce

[21] https://medium.com/analytics-vidhya/hyper-parameters-tuning-practices-learning-rate-batch-size-momentum-and-weight-decay-4b30f3c19ae8

[22] https://optuna.org/

[23] https://www.analyticsvidhya.com/

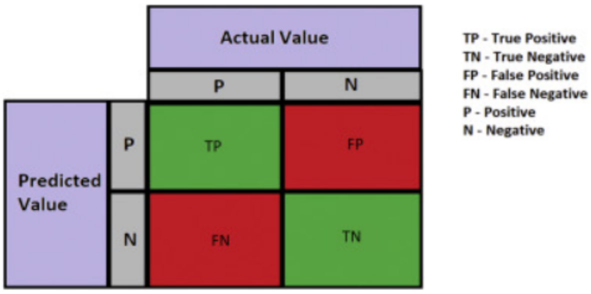


Figure 2: Confusion Matrix

These metrics which are briefly described below can be defined in terms of four attributes viz. True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). Assuming there are two classes i.e. Positive and Negative, TP are those instances of positive reviews which are correctly predicted as positive. TN are those instances which are correctly predicted to be not positive. FP are those which are incorrectly predicted to be positive when they actually are negative. FN are those which are incorrectly predicted to be not negative.

1. Accuracy is the proportion of correctly predicted instances from total instances.
2. Precision is the proportion of correctly predicted positive predictions from total predicted positives (precision = TP / (TP + FP)).
3. Recall is the proportion of the correctly predicted positives from the actual positive predictions (recall = TP / (TP + FN)).
4. F1-Score is the harmonic mean between precision and recall which is useful when there is imbalance between class instances (F1 score = 2 * ((precision * recall) / (precision + recall))).
5. Confusion Matrix[24] is a matrix that describes the performance of the model using attributes like TP, TN, FP and FN. Figure 2 shows the confusion matrix[25] in case of two classes.

3.4.5. Evaluation of Qualitative Metrics

The metrics described in the previous sub section are for quantitative analysis. To analyze the responses of ChatGPT regarding identifying linguistic nuances (labels) present in the texts along with quantitative metrics some qualitative metrics are also discussed which are briefly described below.

[24]https://www.geeksforgeeks.org/confusion-matrix-machine-learning/
[25]https://www.sciencedirect.com/topics/mathematics/confusion-matrix

Correctness. This checks if the linguistic nuance is identified correctly for the given prompt.

Consistency. This verifies if similar output is received for similar input.

Relevance. This checks if ChatGPT can understand context without it being explicitly mentioned.

Robustness. This indicates the understanding of nuance or subtlety by ChatGPT.

3.4.6. Prompts

Seven prompting techniques were used to see how well ChatGPT interpret linguistic nuances present in various sentences. The techniques used along with the prompts can be found in Table 3.

3.5. Misclassification Analysis using LIME

Most of the ML models are blackbox in nature which means their predictions are not easily interpretable. Understanding a model's prediction i.e. why it predicted some label as it did is critical for the model to be reliable [48]. LIME is a technique proposed by [48] and can be used to explain predictions of a classifier in a comprehensible manner. LIME stands for Local Interpretable Model-agnostic Explanation. It is local which means it explains the prediction by learning an interpretable model locally around the prediction. It is model agnostic which means it can be used to explain predictions of any model. Although it can used for explaining any prediction, in this work it was mainly used to elucidate misclassification i.e., those predictions which are incorrect.

Prompting technique	**Prompt**
Zero-shot (no labels)	Analyze the following texts and identify the linguistic nuance each one contains. Answer with following format in a table : Text - linguistic nuance.
Zero-shot	Analyze the following texts and identify which of the following linguistic nuance each one contains - sarcasm, irony, slang, sadness, abbreviation. Answer in following format in a table : Text - linguistic nuance.
One-shot	**Text** : Ugh, I am so annoyed that my boss didn't let me have this weekend off **Linguistic nuance** : Slang Now, analyze the following texts and identify which of the following the linguistic nuance each one contains - sarcasm, irony, slang, sadness, abbreviation. Answer in following format in a table : Text - just linguistic nuance.
Few-shot	**Text** : Woo-hoo! I won the game! **Linguistic nuance** : Slang **Text** : I've been feeling down about work lately. **Linguistic nuance** : Surprise **Text** : Fyi, the meeting is at 2 pm today **Linguistic nuance** : Abbreviation Given the examples, analyze the following texts and identify which of the following the linguistic nuance each one contains - sarcasm, irony, slang, sadness, abbreviation. Answer in following format in a table : Text - just linguistic nuance.
RolePlay(RP)	You are an expert in the English language. Below is a list of linguistic nuances and multiple texts. Analyze each text and identify which of the linguistic nuance it contains. Answer in following format in a table : Text - linguistic nuance. **List of nuances** : sarcasm, irony, slang, sadness, abbreviation.
Chain-of-thought(CoT)	Below is a list of texts. 1. For each text, read it carefully to understand the context and tone. 2. Identify any linguistic nuances present in the sentence. Look for elements such as irony, sarcasm, slang, abbreviation, and sadness. 3. Consider the specific words, phrases, and overall sentence structure that indicate these nuances. 4. Based on your observations, assign a label from one of the following options: irony, sarcasm, slang, abbreviation, sadness. 5. Summarize the results in the form of a table with the following columns: Text and Linguistic Nuance
RP-CoT	You are an expert in the English language. Below is a list of texts. 1. For each text, read it carefully to understand the context and tone. 2. Identify any linguistic nuances present in the sentence. Look for elements such as irony, sarcasm, slang, abbreviation, and sadness. 3. Consider the specific words, phrases, and overall sentence structure that indicate these nuances. 4. Based on your observations, assign a label from one of the following options: irony, sarcasm, slang, abbreviation, sadness. 5. Summarize the results in the form of a table with the following columns: Text and Linguistic Nuance

Table 3: Prompting Techniques Used With Prompts

4. Results

4.1. Exploratory Data Analysis

In this subsection some results of EDA are presented. These can be used to understand the characteristics of the data i.e. textual reviews of the courses.

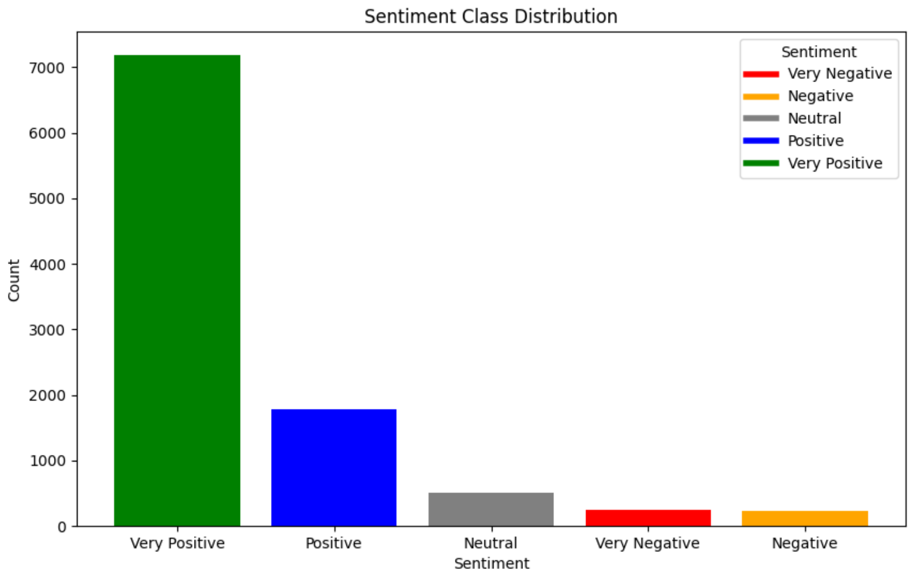


Figure 3: Class Distribution

Figure 3 above shows the distribution of the sentiment labels in the dataset. With a lot of variation in the numbers of text belonging to each class and very large number of texts belonging to class Very Positive, it was clear that the dataset is very imbalanced. Along with train-test method for fine-tuning and validation, to mitigate the imbalance problem, the validation technique of stratified k-fold cross validation was also applied in the process of fine-tuning.

Figure 4 shows the distribution of the length of texts present in the dataset. A vast majority of the texts have less than 100 characters with very few have more than 400 characters. Although LLMs are the most advanced methods to handle large texts it is useful to test their limits.

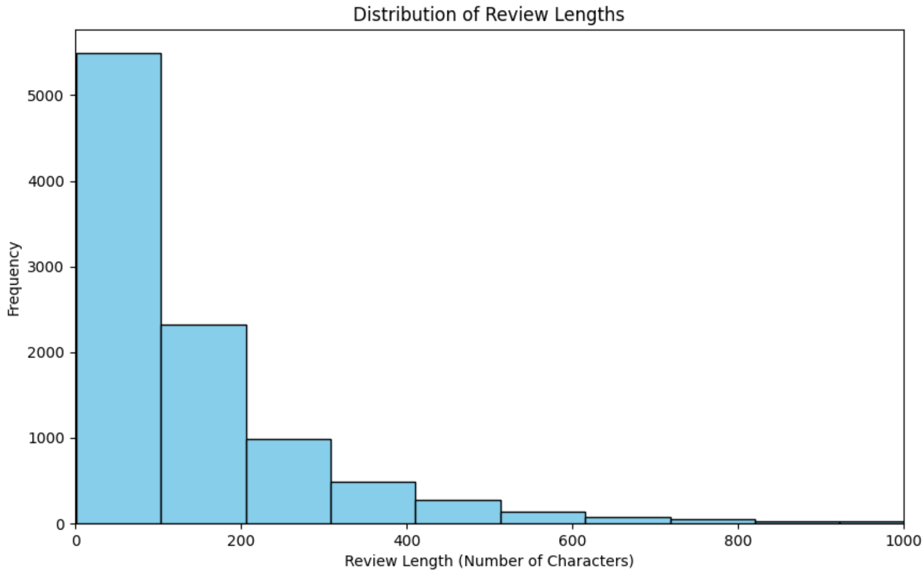


Figure 4: Text Length Distribution.

Figure 5: Negative Wordcloud

Figure 6: Neutral Wordcloud

Figure 7: Positive Wordcloud

Figure 5, Figure 6 and Figure 7 show the wordcloud for the most commonly used words in the texts labeled as Negative, Neutral and Positive respectively. Majority of the words in the negative wordcloud are with negative. However, some positive sounding words are present as they are used in a negative phrase e.g. not good,good but etc. Neutral wordcloud shows both negative and positive words.

4.2. Sentiment Analysis Evaluation

In this sub section the evaluation results of the LLMs is discussed. As mentioned in the methodology, two types of models were considered for comparing with Chat-GPT viz. those that require fine-tuning on the dataset (BERT, XLNet) and those that do not(bart-large-mnli, RoBERTa-large-mnli). These two types are compared for both three labels and five labels classification. For models that require fine-tuning, results of both train-test split and SKF cross validation are presented along with best hyperparameters and performance on unseen data.

4.2.1. Best Hyperparameters

Table 4 shows the best hyperparameters found by Optuna in fine-tuning of both the models. Maximum batch size for XLNet was 16 as using 32 caused memory issues in Kaggle environment. In train-test the split was 80-20 for train and test sets respectively. In K-fold methods, number of splits were 2 and number of trials were 3.

Table 4: Best Hyperparameters of the Fine-Tuned Models with both the Methods

Method \Model	BERT	XLNet
train-test (three labels)	**Learning rate** : 7.75e-05 **Batch Size** : 32 **Epoch** : 2 **Weight Decay** : 1.70e-04	**Learning rate** : 1.26e-05 **Batch Size** : 8 **Epoch** : 2 **Weight Decay** : 8.9e-04
train-test (five labels)	**Learning rate** : 4.79e-05 **Batch Size** : 16 **Epoch** : 3 **Weight Decay** : 5.93e-05	**Learning rate** : 4.54e-05 **Batch Size** : 8 **Epoch** : 2 **Weight Decay** : 3.63e-05
Stratified K-fold (three labels)	**Learning rate** : 3.14e-05 **Batch Size** : 16 **Epoch** : 3 **Weight Decay** : 7.28e-03	**Learning rate** : 1.04e-05 **Batch Size** : 8 **Epoch** : 2 **Weight Decay** : 6.16e-04
Stratified K-fold (five labels)	**Learning rate** : 2.6e-05 **Batch Size** : 8 **Epoch** : 3 **Weight Decay** : 7.78e-04	**Learning rate** : 3.7e-05 **Batch Size** : 16 **Epoch** : 2 **Weight Decay** : 8.46e-06

4.2.2. Models Performance

Table 5: Weighted Evaluation Metrics Of ChatGPT and Models Fine-Tuned using Train-test method with Three Labels

Model\Metric	Accuracy	Precision	Recall	F1-Score
XLNET	**95** %	94 %	**95** %	**94** %
BERT	94 %	93 %	94 %	93 %
ChatGPT	92 %	**95** %	92 %	93 %

Table 6: Weighted Evaluation Metrics Of ChatGPT and Models Fine-Tuned using Train-test method with Five Labels

Model\Metric	Accuracy	Precision	Recall	F1-Score
XLNet	**80** %	**78** %	**80** %	**78** %
BERT	79 %	76 %	79 %	77 %
ChatGPT	57 %	76 %	57 %	61 %

Table 5 shows the evaluation metrics calculated on test part of the dataset for fine-tuned models when train-test method was used and three labels were present. The weighted values of the metrics were calculated due to large imbalance in the data. XLNet showed the best performance in all metrics when fine tuned although the metrics were quite close for all three models and in precision, ChatGPT even surpassed XLNet.

Table 6 shows the evaluation metrics calculated on test part of the dataset for fine-tuned models when train-test method was used and five labels were present. The weighted values of the metrics were calculated due to large imbalance in the data. Here too, XLNet showed the best performance in all metrics when fine tuned doing only slightly better than BERT. ChatGPT's performance though showed a considerable decrease.

Confusion Matrices

Although the evaluation metrics above show the overall performance of each model, confusion matrix helps to learn about models' predictions at individual class level. Confusion matrix for each model is shown in the following figures. The matrix cell values are shown in terms of percentage. This means that each box shows the percentage of predictions matching the criteria e.g. the cell where both true label and predicted label is Neutral will show the percentage of Neutral predictions out of all Neutral predictions. The confusion matrices with actual number of predictions can be found in the Appendix section.

BERT

Figures 8 and Figure 9 show the confusion matrices for fine-tuned BERT when train-test method was used. Darker the shade of blue, higher the number. For instance, the center in the Figure 8 shows that almost 98 % texts were identified correctly in case of three labels and a very high number as Very Positive in case of five labels.

XLNet

Figures 10 and Figure 11 show the confusion matrices for fine-tuned XLNet when train-test was used. Similar to BERT, Positive and Very Positive were the most correctly identified labels in case of three and five labels respectively.

ChatGPT

Figures 12 and Figure 13 show the confusion matrices for ChatGPT predictions on entire dataset. The very dark blue shows that the model predicted the positive and negative sentiment most correctly in case of three labels and negative sentiment (in the bottom right) in case of five labels.

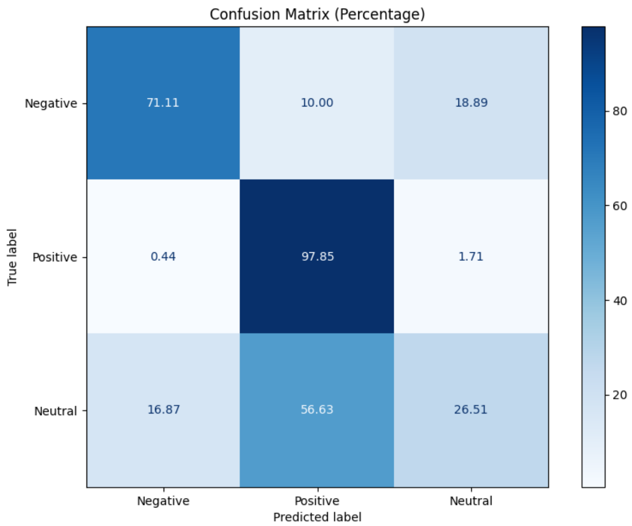


Figure 8: Percentage Confusion Matrix of Fine-tuned BERT using Train-Test in case of Three Labels

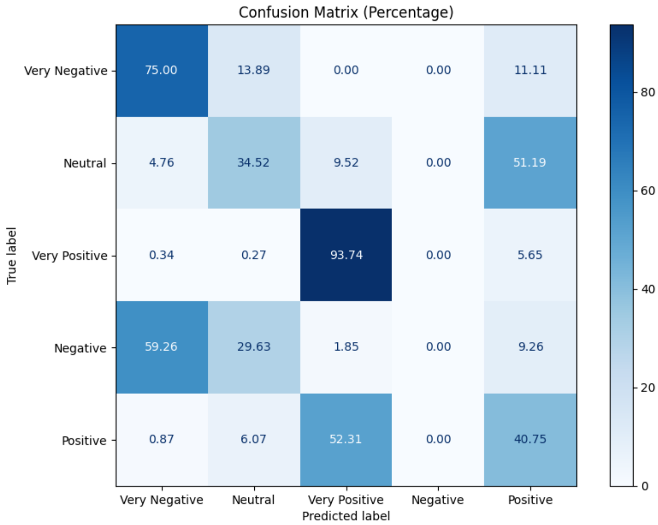


Figure 9: Percentage Confusion Matrix of Fine-tuned BERT using Train-Test in case of Five Labels

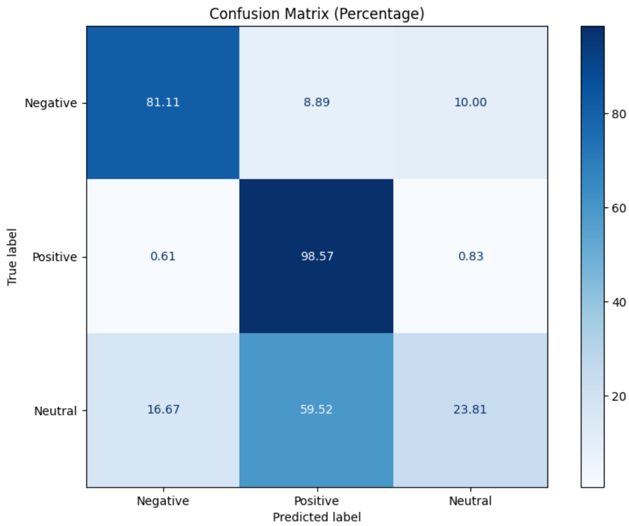


Figure 10: Percentage Confusion Matrix of Fine-tuned XLNet using Train-Test in case of Three Labels

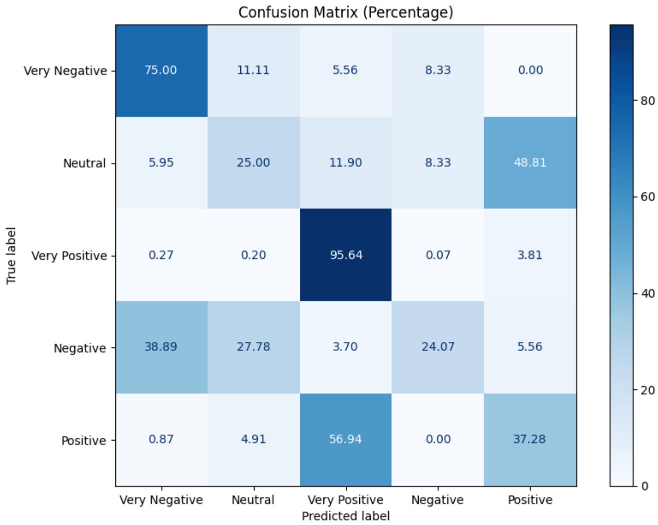


Figure 11: Percentage Confusion Matrix of Fine-tuned XLNet using Train-Test in case of Five Labels

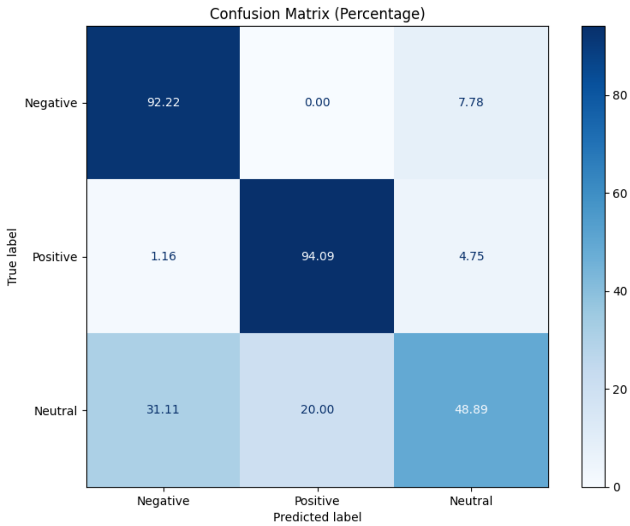


Figure 12: Percentage Confusion Matrix of ChatGPT when Test Split Dataset was used in case of Three Labels

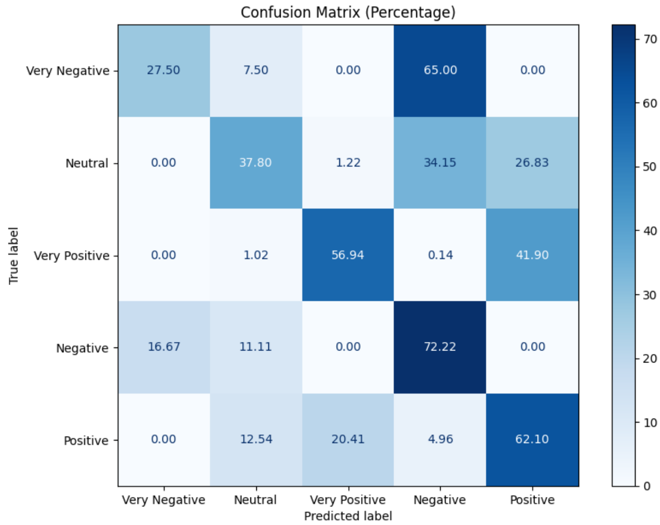


Figure 13: Percentage Confusion Matrix of ChatGPT when Test Split Dataset was used in case of Five Labels

4.2.3. Stratified K-fold Cross Validation

Table 7: Weighted Evaluation Metrics Of ChatGPT and Fine-Tuned Models with Three Labels

Model\Metric	**Accuracy**	**Precision**	**Recall**	**F1-Score**
BERT	**98** %	**98** %	**98** %	**98** %
XLNet	95 %	94 %	95 %	95 %
ChatGPT	90 %	93 %	90 %	91 %

Table 7 shows the evaluation metrics calculated for fine-tuned models when stratified k-fold cross validation was used and three labels were present. The weighted values of the metrics were calculated due to large imbalance in the data. BERT showed the best performance in all metrics when fine tuned.

Table 8 shows the evaluation metrics calculated for fine-tuned models when stratified k-fold cross validation was used and five labels were present. The weighted values of the metrics were calculated due to large imbalance in the data. BERT once again showed the best performance in all metrics when fine tuned.

Table 8: Weighted Evaluation Metrics of ChatGPT and Fine-Tuned Models with Five Labels

Model\Metric	Accuracy	Precision	Recall	F1-Score
BERT	**90** %	**90** %	**90** %	**90** %
XLNet	80 %	78 %	80 %	78 %
ChatGPT	56 %	75 %	56 %	61 %

Confusion Matrices

Similar to train-test method, confusion matrices for each model are shown in the following figures.

BERT

Figures 14 and Figure 15 show the confusion matrices for fine-tuned BERT when SKF was used. Darker the shade of blue, higher the number. For instance, the bottom right corner in Figure 14 shows that almost 100 % Positive texts were identified correctly in case of three labels and as Very Positive in case of five labels.

XLNet

Figures 16 and Figure 17 show the confusion matrices for fine-tuned XLNet when SKF was used. Similar to BERT, Positive and Very Positive were the most correctly identified labels in case of three and five labels respectively.

ChatGPT

Figures 18 and Figure 19 show the confusion matrices for ChatGPT predictions on entire dataset. The very dark blue in the bottom right shows that the model predicted the positive sentiment most correctly in case of three labels and negative sentiment in case of five labels.

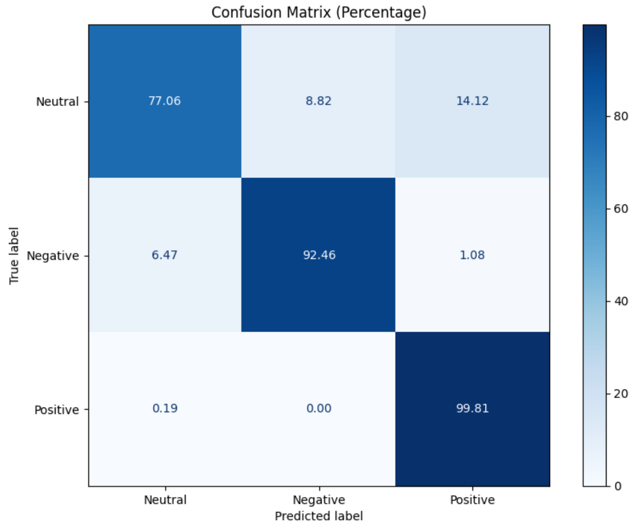


Figure 14: Percentage Confusion Matrix of Fine-tuned BERT using SKF in case of Three Labels

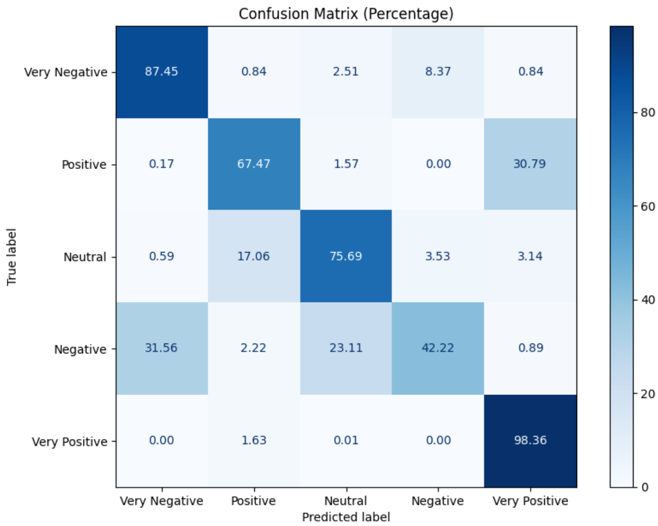


Figure 15: Percentage Confusion Matrix of Fine-tuned BERT using SKF in case of Five Labels

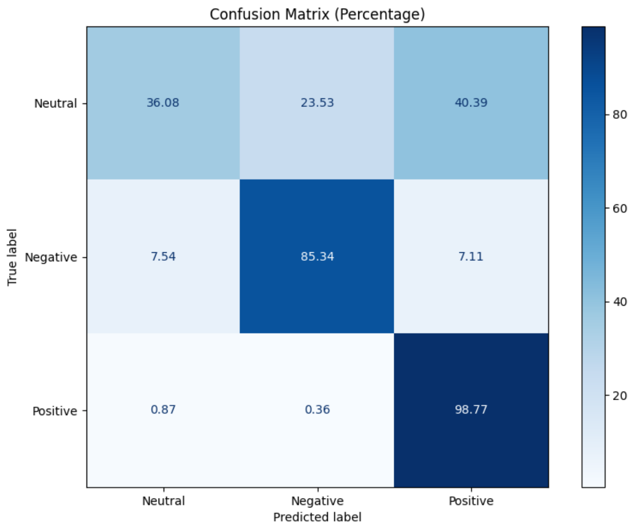


Figure 16: Percentage Confusion Matrix of Fine-tuned XLNet using SKF in case of Three Labels

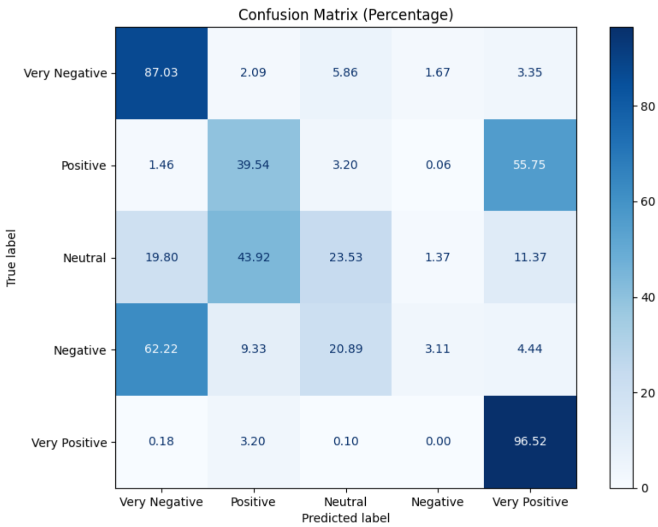


Figure 17: Percentage Confusion Matrix of Fine-tuned XLNet using SKF in case of Five Labels

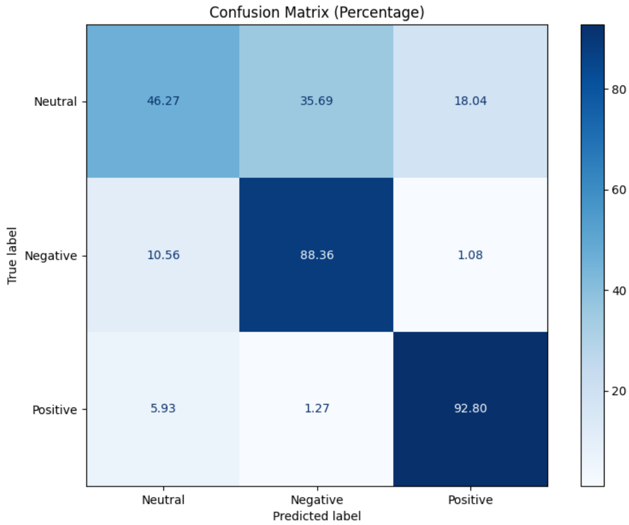


Figure 18: Percentage Confusion Matrix of ChatGPT when entire dataset was used in case of Three Labels

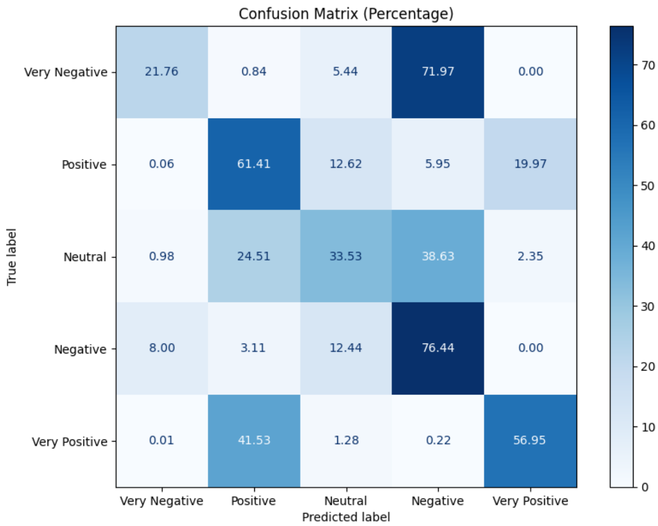


Figure 19: Percentage Confusion Matrix of ChatGPT when entire dataset was used in case of Five Labels

Table 9: Weighted Evaluation Metrics of ChatGPT and already Fine-Tuned Models used with Three Labels

Model\Metric	Accuracy	Precision	Recall	F1-Score
ChatGPT	**90** %	**94** %	**90** %	**92** %
RoBERTa-large-mnli	89 %	91 %	89 %	89 %
bart-large-mnli	89 %	90 %	89 %	89 %

4.2.4. Fine-Tuned Models

Table 9 shows the evaluation metrics calculated for those models which were not fine-tuned on our dataset but were already fine-tuned for sentiment analysis on other data and suitable for three labels. The weighted values of the metrics were calculated due to large imbalance in the data. Both variations of RoBERTa and BART showed almost the same performance whereas ChatGPT was better than both on all four metrics.

Table 10: Weighted Evaluation Metrics of ChatGPT and already Fine-Tuned Models used with Five Labels

Model\Metric	Accuracy	Precision	Recall	F1-Score
bart-large-mnli	**71** %	68 %	**71** %	**69** %
ChatGPT	56 %	**75** %	56 %	61 %
RoBERTA-large-mnli	34 %	69 %	34 %	37 %

Table 10 shows the evaluation metrics calculated for those models which were not fine-tuned on our dataset but were already fine-tuned for sentiment analysis on other data and suitable for multi class scenario. The weighted values of the metrics were calculated due to large imbalance in the data. bart-large-mnli showed the better performance than ChatGPT on three metrics while RoBERTa based model showed significantly bad performance compared to other two on all metrics.

bart-large-mnli

Figures 20 and Figure 21 show the confusion matrices for bart-large-mnli. The model shows performed poorly to predict Neutral sentiment in case of both three and five labels.

RoBERTa-large-mnli

Figures 22 and Figure 23 show the confusion matrices for RoBERTa-large-mnli. In case of three labels it shows a very good performance in detecting Positive labels.

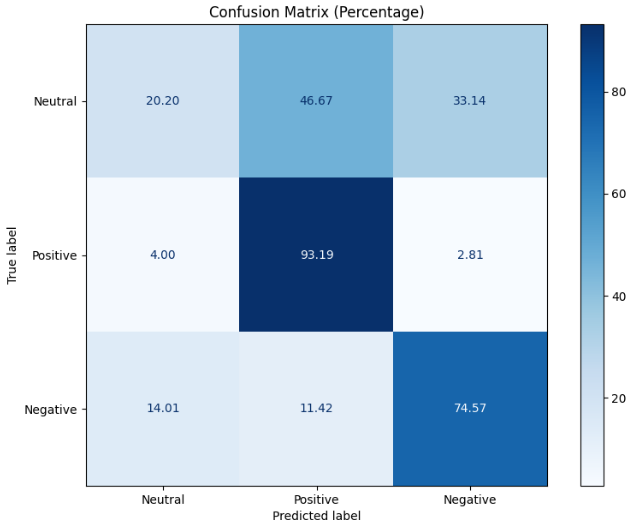


Figure 20: Percentage Confusion Matrix of bart-large-mnli in case of Three Labels

However, in case of five labels it shows varied performance.

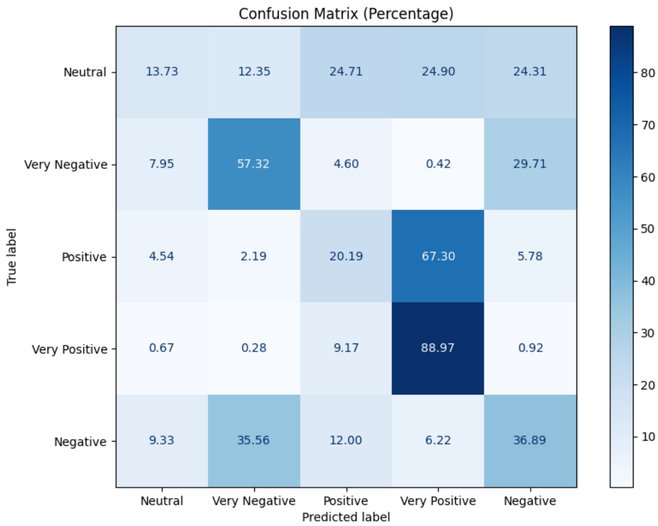


Figure 21: Percentage Confusion Matrix of bart-large-mnli in case of Five Labels

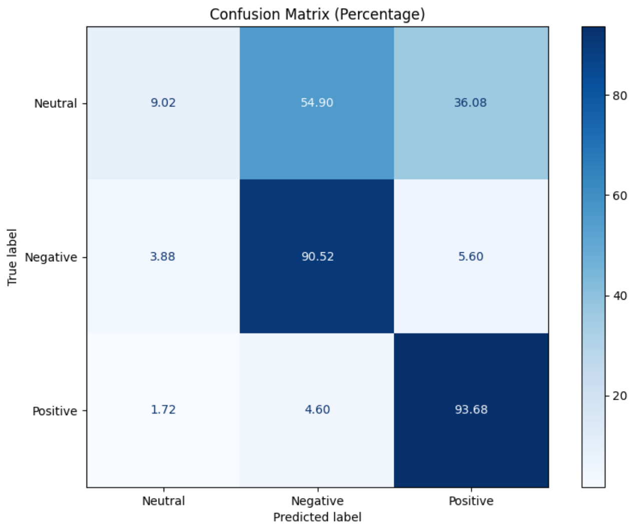


Figure 22: Percentage Confusion Matrix of RoBERTa-large-mnli in case of Three Labels

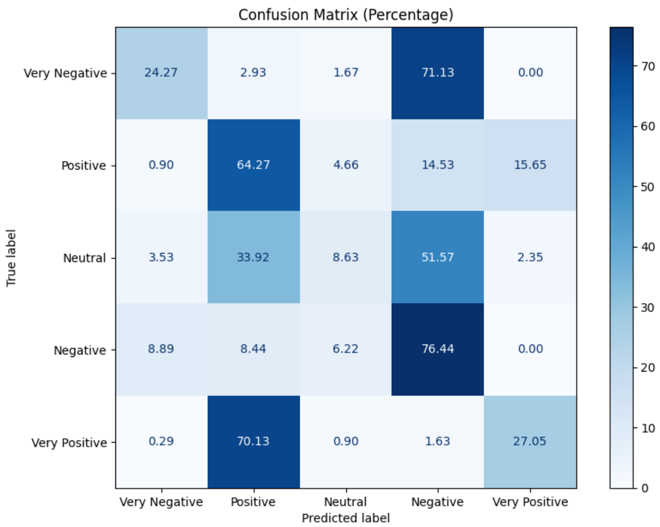


Figure 23: Percentage Confusion Matrix of RoBERTa-large-mnli in case of Five Labels

All models viz. BERT, XLNet, ChatGPT, bart-large-mnli and RoBERTa-large-mnli have shown the lowest performance for Neutral labels in case of three labels. XLNet predicted more than half the texts as Very Positive when the actual label was Positive in both train-test and SKF. These figure is 67 % for bart. More than half the texts are predicted as Very Negative and Very Positive by BERT when the actual labels are Negative and Positive respectively in the test part of the split dataset. In case of RoBERTa and ChatGPT, 70 % and 56 % of the Very Positive true labels are predicted as Positive on the whole dataset. On more than 70 % of the instances, Very Negative are predicted as Negative by ChatGPT. This figure is 65 % on the test part of the split dataset for ChatGPT. These key highlights strongly indicate that all these models struggle to distinguish between either Positive-Very Positive or Negative-Very Negative.

4.2.5. Evaluation on Unseen Data

The evaluation metrics discussed in previous subsection were based on the original dataset. Testing the models on unseen data i.e. the data was not part of their train or evaluation splits gives an idea about how it might perform on new data.

Table 11: Weighted Evaluation Metrics of ChatGPT and Fine-Tuned Models with Three Labels on Unseen Data

Model\Metric	**Accuracy**	**Precision**	**Recall**	**F1-Score**
ChatGPT	**78** %	**77** %	**78** %	**77** %
XLNet	70 %	68 %	70 %	68 %
BERT	68 %	71 %	68 %	67 %

Table 12: Weighted Evaluation Metrics of ChatGPT and Fine-Tuned Models with Five Labels on Unseen Data

Model\Metric	**Accuracy**	**Precision**	**Recall**	**F1-Score**
ChatGPT	**58** %	**74** %	**59** %	**65** %
BERT	45 %	45 %	45 %	43 %
XLNet	43 %	42 %	43 %	36 %

Tables 11 and 12 show the evaluation metrics calculated for those models which were fine-tuned on our dataset after testing on unseen data. The unseen dataset was nearly balanced. ChatGPT showed better performance than both the other models in case of three as well as five labels. The metrics of BERT and XLNet declined drastically on unseen data hinting at overfitting.

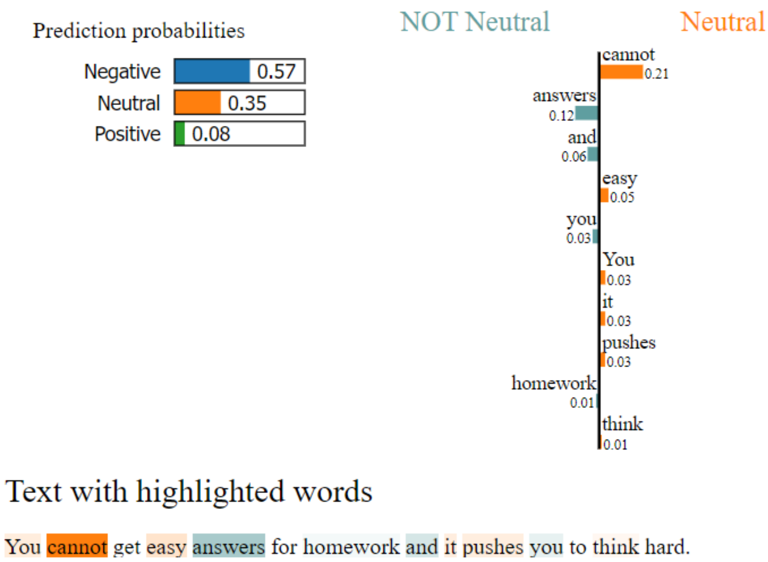


Figure 24: BERT Misclassification - Positive text predicted as Negative (Three Labels)

4.2.6. Misclassification Analysis

Each model has a different rational to its predictions and there are a number of predictions which are not matching with the actual label. It's useful to analyze the incorrect predictions using the tool like LIME as described in the methodology. Some examples of wrong predictions (from both three labels and five labels) are explained visually for BERT and XLNet and presented in this sub section. For incorrect predictions of ChatGPT, it was prompted to explain its reasoning and that will also be discussed here.

BERT

Figure 24 shows an example of a text which is Positive. However, the model predicted is to be Negative. The colors in the visual are significant. The colors assigned to each word indicate how much influence it had on the predictions. The darker shade means stronger influence. In the above mentioned example, there are words on either side with varying degree of influence. However, the word "cannot" was the most influencing word which resulted in model considering the text to be negative. However, after reading the text it can be understood that it's not a negative statement. Thus, the model got it wrong here.

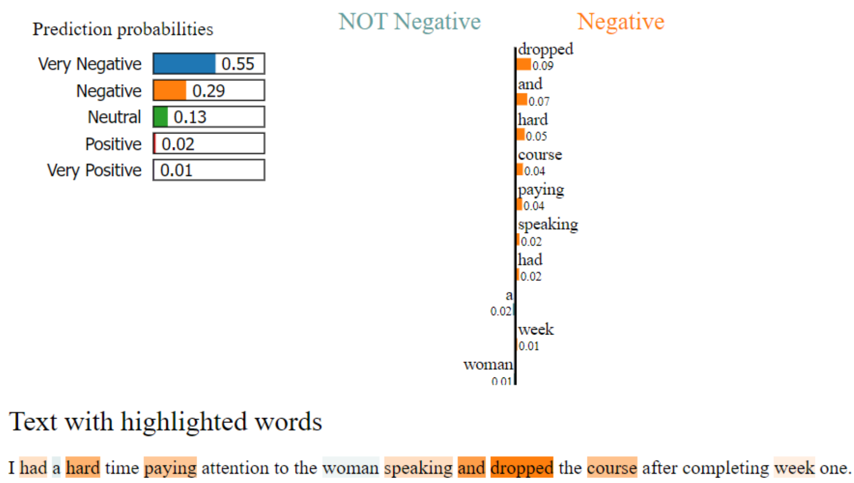


Figure 25: BERT Misclassification - Negative vs Very Negative (Five Labels)

Figure 25 shows an example of a text which has Negative as the actual label. However, there are words indicating a stronger sentiment. The word "dropped" indicates a very strong sentiment as leaving the course could be considered an extremely negative review. Here, the model got it right and the actual label may not be most accurate.

XLNet

Figure 26 shows an example of a text which is Neutral. However, the model predicted is to be Positive. Moreover, it is strongly on the side of positive with the probability of 97% which is clearly not the case. The words "Nice" and "Clear" in darker shades outweigh other words. However, the second part of the text indicates that it is not entirely positive review. Thus, the model got it wrong.

Figure 27 shows an example of a text which is predicted as Very Negative but has actual label as Negative. Emphasis on the word tedious makes this text Very Negative as per the model. This text is certainly Negative. However, whether it is Very Negative is ambiguous with varying degree of negativity of influencing words.

ChatGPT

Figure 28 shows an illustration of misclassifiation of ChatGPT. Here, ChatGPT was asked to explain why it assigned the label Negative to the text and in response it identifies both positive and negative aspects from the text and yet labels it as negative whereas actually this is a neutral text.

Figure 29 shows another illustration of misclassifiation of ChatGPT. Here, when

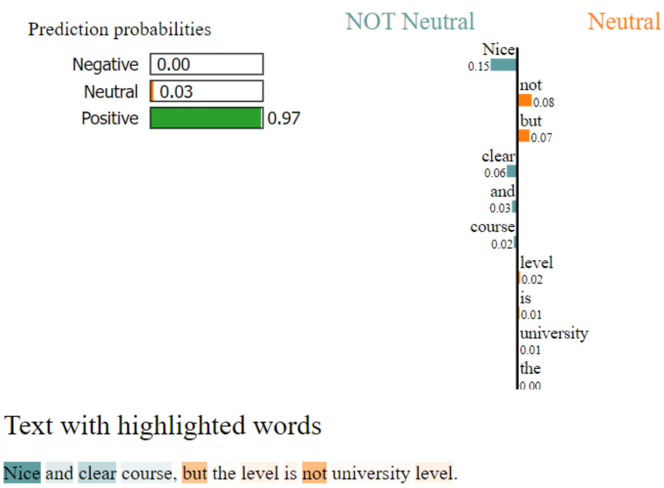


Figure 26: XLNet Misclassification - Neutral as Positive (Three Labels)

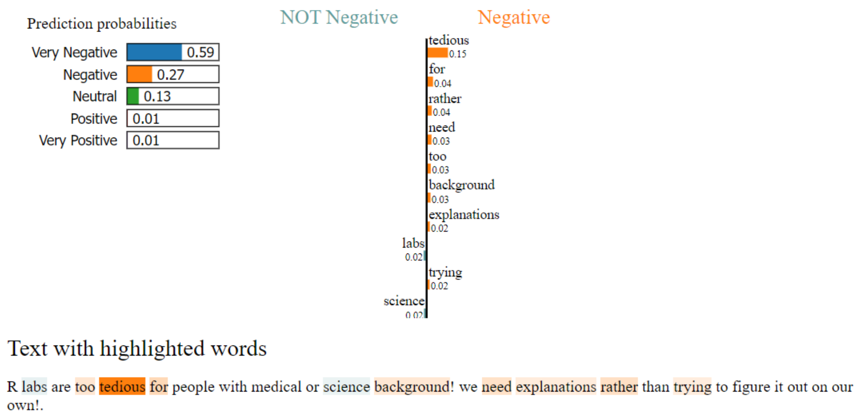


Figure 27: XLNet Misclassification - Negative vs Very Negative (Five Labels)

Review	Sentiment
Good Angular exercises. But I think that testing should get its own segment. Becomes a bit lengthy with 20 min videos and only scraping the surface on the subject.	Negative

Explanation:

The review expresses Negative sentiment primarily due to the following phrases:

- "But I think that testing should get its own segment."
- "Becomes a bit lengthy with 20 min videos and only scraping the surface on the subject."

These phrases indicate dissatisfaction and critique towards the course structure and the depth of coverage on testing, suggesting room for improvement. Therefore, despite the positive remark about the Angular exercises, the overall sentiment leans towards Negative.

Figure 28: ChatGPT Misclassification - Neutral as Negative (Three Labels)

Text	Sentiment
79 USD? just to be able to submit peer-graded assignments... talk about greed	Negative

Explanation: The text expresses a negative sentiment regarding the cost of 79 USD for submitting peer-graded assignments. The term "greed" indicates strong disapproval and criticism towards what the reviewer perceives as an excessive or unfair pricing practice.

Figure 29: ChatGPT Misclassification - Negative vs Very Negative (Five labels)

ChatGPT was asked to explain why it assigned the label Negative to the text, it correctly identifies the tone of the statement. The word "greed" has a strongly negative connotation. ChatGPT identifies the strong disapproval and criticism towards the review in terms of price. Yet, it assigns the Negative label to the text while its actual label is Very Negative.

Thus, there are some points which can be noted from the misclassification which can also be confirmed by the confusion matrix. There are many examples where the texts with actual labels as very positive and very negative are predicted to be positive and negative respectively. This indicates that the models have some difficulty in differentiating between these two categories of labels. Additionally, actual labels of some texts are incorrect and there are multiple instances where they are ambiguous e.g. they could be called positive or very positive or vice versa.

4.3. Linguistic Nuances Evaluation

There are four parameters on which ChatGPT was evaluated qualitatively as mentioned in the methodology section. Assuming the linguistic nuances as labels, quantitative metrics of the predictions of each technique were also calculated. which are shown in the table 13.

Along with these metrics confusion matrix for each technique was also created which can help in learning predictions of individual labels in more details and also in further qualitative findings.

Zero-shot without labels is not a formal prompting technique. It was added to see the response from ChatGPT. Although all its evaluation metrics are 0 it did well in identifying the jist of the text. Among the other techniques CoT turned out to be the most effective overall.

Table 13: Weighted Evaluation Metrics of Prompting Techniques

Technique\Metric	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Zero-shot (no labels)	0 %	0 %	0 %	0 %
Zero-shot	82 %	86 %	82 %	82 %
One-Shot	80 %	93 %	80 %	84 %
Few-Shot	66 %	**95** %	66 %	77 %
RP	75 %	**95** %	75 %	81 %
CoT	**88** %	**95** %	**88** %	**90** %
RP-CoT	81 %	94 %	81 %	85 %

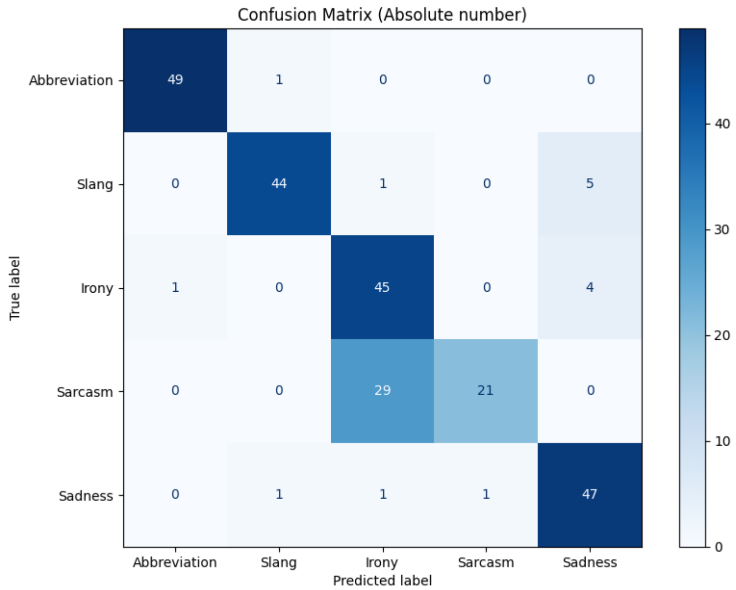


Figure 30: Confusion Matrix of Zero-shot

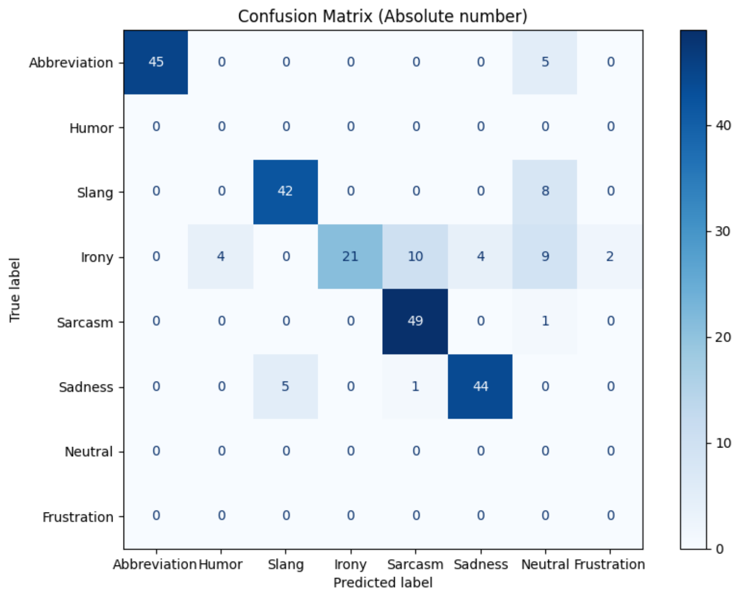


Figure 31: Confusion Matrix of One-shot

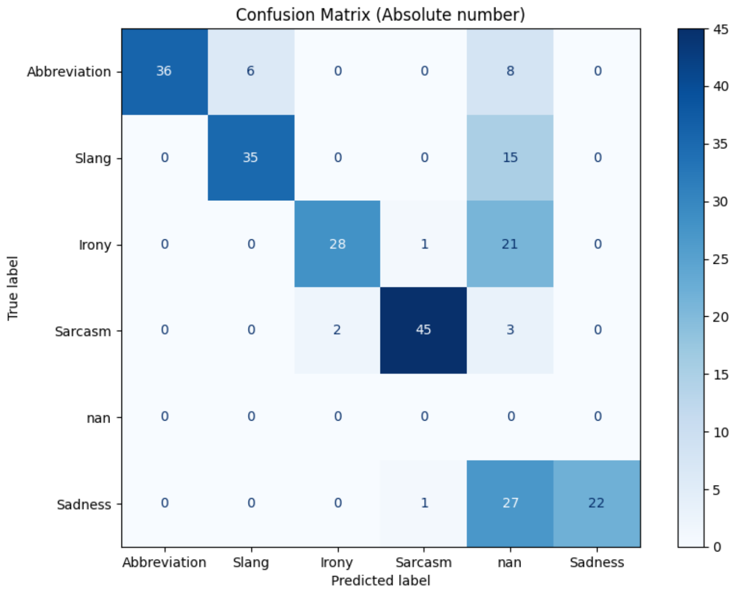


Figure 32: Confusion Matrix of Few-shot

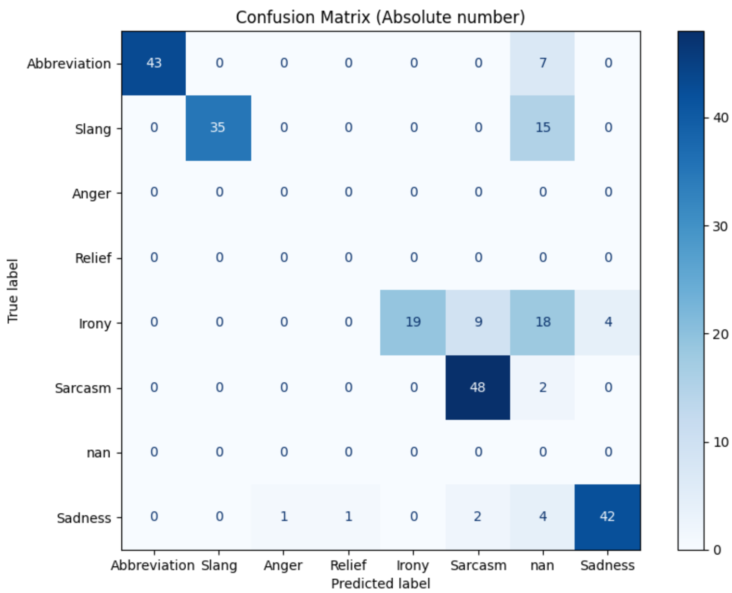


Figure 33: Confusion Matrix of RolePlay

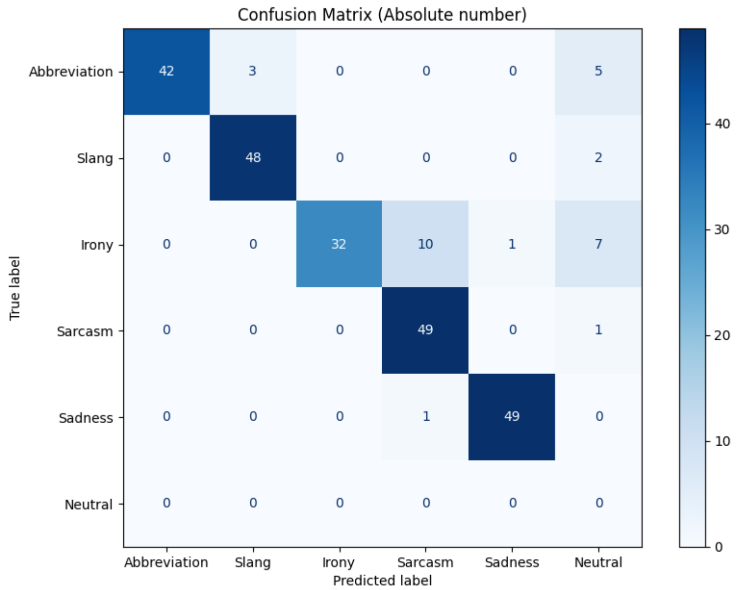


Figure 34: Confusion Matrix of CoT

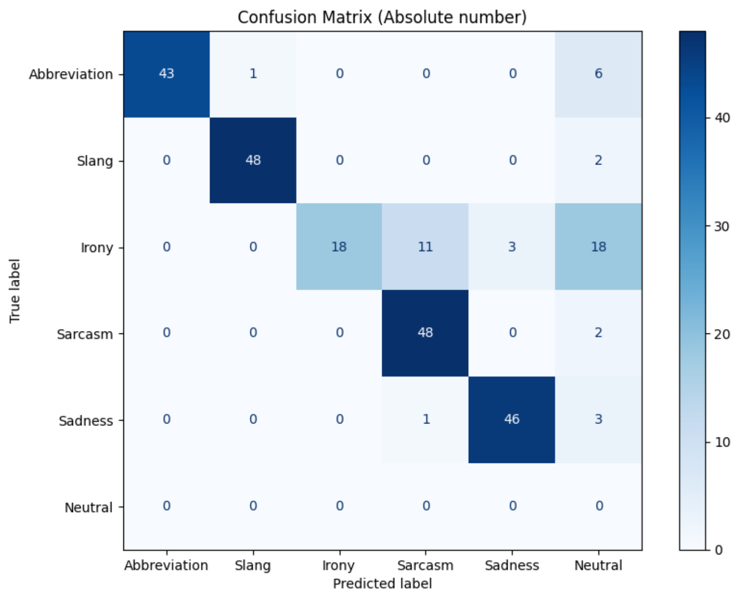


Figure 35: Confusion Matrix of RP-CoT

Figures 30, 31, 32, 33, 34, 35 show the confusion matrices of the predictions of linguistic nuances for each prompting technique. The values shown in the matrices are absolute numbers of predictions as opposed to percentage.

Azure portal for IE doesn't recognize my login...but access it through Chrome I get right in	Criticism of software compatibility issues.

Figure 36: Zero-shot without label - Inconsistency

Azure portal for IE doesn't recognize my login...but access it through Chrome I get right in	Irony

Figure 37: Zero-shot - Inconsistency

Azure portal for IE doesn't recognize my login...but access it through Chrome I get right in	Frustration

Figure 38: One-shot - Inconsistency

Azure portal for IE doesn't recognize my login...but access it through Chrome I get right in	None

Figure 39: Few-shot - Inconsistency

Correctness. This metric was used to check if and to what extent the nuances are identified correctly by ChatGPT. The accuracy metric can be a good indicator of this. Among the prompting techniques, the most accurate responses were received to a CoT prompt. Among the linguistic nuances Sarcasm was the most accurately identified nuance. Sarcasm is considered to be a nuance which is not straightforward. Irony, on the other hand is the nuance that it struggled to identify the most. It is the only nuance which was identified correctly on less than 50 % of the occasions.

Consistency. This was evaluated at two levels. First, it was checked between the techniques i.e. if the same nuance is identified when each of the technique was used from any given text. Second, within the same technique, does it identify the nuances consistently on different texts. CoT showed the most consistent responses. For the latter, the responses were inconsistent. For instance, when the text were provided in batches with a batch of 20 all containing the same nuance, the responses were varied. Figures 36, 37, 38, 39 show an example of inconsistency. When same text was provided using four different techniques in different chats it returned four different responses. The text is shown in the box on the left hand side and response on the right.

I'm doing an AMA on Reddit tonight.	Announcement of an online interactive event.

Figure 40: ChatGPT's Response when no label was mentioned - Example

I'm so glad millionaire families are getting a little extra at the expense of the working poor!	Sarcasm about economic disparities.

Figure 41: ChatGPT's Response when no label was mentioned - Example

I have personally experienced this gut-wrenching feeling and kicked myself later for making those dumb mistakes that result when anxiety gets in the way.	Intense regret and self-reproach linked to anxiety-induced errors.

Figure 42: ChatGPT's Response when no label was mentioned - Example

Relevance. This was analyzed to test if ChatGPT can understand the context even when it wasn't mentioned explicitly in the text. Interestingly, Zero-shot without labels showed very good responses where for each text it described what was the text about. Figures 40, 41, 42 show some of the example responses. The text is in the left column while the response in the right. In all three examples it it sensed the message that was conveyed through the texts and mentioned it.

Robustness. This metric assessed the understanding of the nuance or subtleties in the text. Some texts didn't contain subtlety. For instance, some of those containing abbreviations. But among those texts which contained them, vast majority of the times, ChatGPT could understand it. In some cases, where it predicted the label which wasn't a part of the list that it was asked to choose from, it wasn't completely wrong as some of those texts contained more than one emotion or nuance. Overall, barring the cases of irony and some examples from other nuances, ChatGPT did reasonably good.

Overall, ChatGPT showed some randomness by making up labels on its own on a few examples. It struggled in detecting Irony in a text. Barring these, the overall result from the examples show a promising ability of ChatGPT.

5. Discussion

The main objectives of this thesis were to conduct a literature review of the SOTA methods for SA, compare the performance of ChatGPT with some of the other SOTA models (BERT, XLNet, bart-large-mnli, RoBERTa-large-mnli) on the task of SA and explore prompt engineering and linguistic nuance recognition in SA using ChatGPT. In this section some of the key findings, limitations of the work and ethical considerations are described.

5.1. Key Findings

Firstly, there are various SOTA methods used for SA and five of them are discussed in this work with each having its own features and weaknesses viz. Ensemble Learning, Transfer Learning, Graph Neural Network, Multimodal Sentiment Analysis and Large Language Models. Some of these methods e.g. Ensemble learning were introduced a long time ago. However, they have been used with SOTA models like pre-trained models. Transformer based methods are the most advanced of all the methods with their superior ability to understand context even in long texts. Transformers are the architecture that help the models focus on the most important parts of the text built on the concept of Attention which was proposed by researchers at Google. LLMs are large as they are trained on massive amount of data which gives them an exceptional ability in NLP tasks. There are methods for diverse types of data. For instance, GNN is useful when the data can be represented using a graph with nodes and edges like social media platforms where people are nodes and their connections are edges. There are methods to analyze sentiments from multiple modalities like texts, videos and images. These can be useful for a deeper analysis of sentiments and suitable for a lot of data on the web which is available through platforms like video streaming. There are studies highlighting superior performance of all these methods. However, all of these are computationally very expensive with hardware requirements of GPUs and even more advanced processors.

There are some observations about SA. Firstly, EDA clearly shows that the data is highly imbalanced with much higher number of samples belonging to a single class. However, this was kept as it is as opposed to taking the same number of samples from each class. This was done to test the models on real dataset as finding a real dataset which is balanced is not a certainty in real scenarios. Previous studies have shown that ChatGPT demonstrates a good performance in Zero-shot SA tasks. In simple sentence level SC as Positive or Negative it showed the accuracy of 93.12 % [59] and 93.6 % [66] on the SST-2 [50] dataset. In this thesis, ChatGPT showed the accuracy of 92 % in SA task when there are three labels. In complex SA tasks it trails behind fine-tuned models like BERT in some cases and is comparable in some other cases [66] [59]. In case of three labels (Positive, Negative, Neutral) ChatGPT's performance is comparable to fine-tuned BERT, XLNet as the comparison with SST-2 shows and better compared to models like bart-large-mnli and RoBERTa-large-mnli which have been fine-tuned and are suitable for classification tasks. In case of five labels (Very Negative, Negative, Neutral, Positive, Very Positive) all the models showed decrease in performance. ChatGPT's performance declined more than the other models. [59] tested E2E-ABSA (End-to-End Aspect Based Sentiment Analysis) and observed the accuracy of 77.75 % and 69.14 % of accuracy by fine-tuned BERT and ChatGPT respectively on 14-Restaurant [44] dataset. In the same study the numbers were 66.05 % and 49.11 % respectively on the 14-Laptop [44] dataset. Although SA with five labels is not equivalent of E2E-ABSA, it is a complex SA task and thus is considered for comparison. In this thesis work, considering the class imbalance, if F1-scores are compared between models in both methods in five labels, the difference between highest performing fine-tuned BERT and ChatGPT is 28 % and 16 % in SKF and train-test respectively. Thus, the results in this study are similar to previously observed.

I have officially hit the halfway point of the course and am three weeks ahead of schedule. The last time I took a Calculus class was 17 years ago, during my Senior year of high school. It was the first time I ever got a B in Math, and ruined my love for the subject. Since then, I had all but abandoned my love for the subject because I thought I had a learning deficiency for Calculus. It turns out that my teacher was just sub-par, or leagues below Dr. Fowler. As someone who is strong with numbers, he is the perfect professor to bring out the conceptual learning student in me. The most impressive thing though is how he has delivered rules to me in an almost photographic format. After each lesson, which are all theoretical then provide graphical and numerical examples, I do one practice problem with what I absorbed and check the answer. It is normally incorrect, but after reading the correct answer and explanation I fly through the next five. Although I am sure that not having closed book graded quizzes and more practice problems may subtract from his classroom course a bit, this is an amazing confidence booster and has propelled me back on track!	Very Positive

Figure 43: ChatGPT's Prediction For a Long Text

The distinction between Positive-Very Positive and Negative-Very Negative is where all the models struggled and that lead to the decline in performance. Some examples of this are shown in the subsection of Misclassification analysis. To mitigate the issue of class imbalance Stratified K-fold cross validation techniques was used and its results compared to the results of train-test approach are not massively different with metrics like high F1-score remaining similar. Fine-tuned BERT and XLNet showed overfitting as their performance declined drastically on unseen data.

There were some texts which were extremely long with more than a hundred words. The models were able to process these with correct predictions on many occasion confirming their superior ability over traditional methods. Figure 43 shows an example of a long review of a course and ChatGPT's prediction of Very Positive on it which matches with the actual label. When it was asked to explain its prediction it returned a detailed response which can be seen in the Figure 44. There are multiple aspects that the reviewer has talked about in the text. His sentiment towards

The review is labeled as "Very Positive" due to several factors that highlight the transformative impact the course, particularly Dr. Fowler's teaching style, has had on the student. Here's a breakdown of how the analysis led to this rating:

1. **Personal Transformation**: The reviewer reflects on a past negative experience with calculus that led them to lose interest in the subject. The positive change in their perspective after taking this course signifies a significant transformation, often seen as a strong positive endorsement.
2. **Effective Teaching**: The reviewer praises Dr. Fowler's ability to communicate complex concepts clearly, which they describe as being delivered in an "almost photographic format". This highlights the effectiveness of the teaching method, making the course not only comprehensible but also engaging.
3. **Enhanced Understanding and Confidence**: The reviewer mentions that despite initially getting answers wrong, the explanations provided enable them to quickly correct mistakes and successfully tackle subsequent problems. This improvement in understanding and skill is a testament to the course's effectiveness in enhancing student capability and confidence.
4. **Progress and Pace**: The reviewer notes that they are three weeks ahead of schedule, indicating that the course content and the instructor's delivery have motivated them to engage more deeply and progress faster than expected.
5. **Comparison with Past Experience**: There is a comparison between past instructors and Dr. Fowler, with Dr. Fowler being highly favored. This comparative appreciation further emphasizes the positive impact of the current course experience.

These elements collectively contribute to a very positive learning experience, meriting a "Very Positive" label. The review clearly indicates a high level of satisfaction and a renewed passion for calculus, driven by effective instruction and a supportive learning environment.

Figure 44: ChatGPT's Explanation On Prediction For a Long Text

the course is indicted after some background is provided. Yet, ChatGPT is able to collectively analyze every part and predict. Despite this observation, it also worth noting that there is a limit to the length of a single message sent to ChatGPT as well as the length of a single conversation with it.

I made myself a vegan dinner treat and now it's time to catch up with Hannibal season 3!	Ironic juxtaposition of veganism with a show about a cannibal.

Figure 45: ChatGPT's Response when no label was mentioned - Example

I feel a flare of anger because it still pains me to think of Mal being abused like that, but I can't help wonder now if he might be right.	Mixed emotions of anger, pain, and doubt regarding a situation involving abuse.

Figure 46: ChatGPT's Response when no label was mentioned - Example

What's your OOTD? I don't know which shoes to wear!	Asking about daily fashion choice, using abbreviation for "outfit of the day".
FWIW, I think it's going to rain anyway.	Providing an opinion that might be useful, using abbreviation for "for what it's worth".
Ngl, I'm ready for Friday	Expressing anticipation for the weekend, using abbreviation for "not gonna lie".

Figure 47: ChatGPT's Response when no label was mentioned - Example

There are multiple techniques used to instruct ChatGPT called prompting techniques to get desired results. Six of them (plus one tweaked) were used in this work viz. Zero-shot, One-shot, Few-shot, RP, CoT, RP-CoT. Zero-shot, One-shot and Few-shot differ on the number of reference examples provided to ChatGPT before asking it to perform the actual task with 0, 1 and few respectively. In this work the number of examples used don't make a significant difference to the outcome of identifying linguistic nuance present in the texts. The techniques were evaluated using four criteria of Correctness, Consistency, Relevance and Robustness. CoT, where step by step instructions are given for the task proved to be the most effective technique. Sarcasm, irony, abbreviations are considered difficult linguistic nuances for ChatGPT [26]. However, it was able to detect sarcasm and abbreviations reasonably well albeit the result was not perfect. Irony, on the other hand is something that the model failed to identify on many samples.

Zero-shot without labels, an experimental tweak helped to get interesting observations about its linguistic ability. Some examples of this are can be seen in the figures above. Figure 45 shows an example where ChatGPT showed that it knows juxtaposition, veganism. Moreover, even though there is a real show called Hannibal, it correctly identified the term to be related to cannibal which is shown on the right side. Right side of the Figure 46 shows that it could detect anger, pain and doubt expressed in the single text where the actual label was sadness. Figure 47 demonstrate that it not only knows modern abbreviations used in conversations but also the situations where they are used without the context being provided. While there were no labels, ChatGPT demonstrated impressive knowledge of the English language with diversity of the words it used to precisely express descriptive prediction.

5.2. Limitations

Despite the key findings of this work there are certain limitations and challenges about ChatGPT and this thesis work that need to be discussed in order to analyze the results in a broader view.

Dataset. The unequal distribution of the class labels in the dataset could have affected the performance of the models which were fine-tuned before being used for SA. However, the validation method used for mitigating that didn't suggest it concretely. Moreover, ChatGPT was used via the web interface with no fine-tuning involved and there was an assumption that it will treat each text independently.

Prompts sensitivity and samples. The number of texts used for identifying linguistic nuances using different prompting techniques was kept lower. This was done keeping the time constraints required for the analysis as for each text was tested with seven prompting techniques and consequently analyzed. Trying with more texts might reveal new or different findings which could be also useful to generalize the observations for more predictibility of the models and prompts. The prompts designed in this work were designed by following the basics of each technique. However, as ChatGPT has shown different prompts may produce different outcomes [56] even on the same text which could pose challenges to reproduce the results as well.

Language. All the texts used in this work for both SA and prompt engineering for linguistics nuances were in English. The results could have been different if the same process was followed on the texts in some other language especially those which are low resource languages [4]. However, the lack of sufficient resources like standard datasets is a limitation as it is for those languages. Different variations of English language used in different regions of the world could produce some more insights. Particularly, the slang samples used in this work are mainly used in American English.

Bias. LLMs in spite of their superior capabilities are believe to contain biases [26]. The models that are trained on massive amount of data on the internet like ChatGPT (GPT) might have affected their predictions as they would have learned about prejudices, biases present in the data from different data sources [47]. This is a hard issue that will require training the models on diverse datasets for adding more fairness to the predictions. Besides, the blackbox nature of the predictions means that they are not easy to interpret.

Cost. The transformer based models are extremely resource intensive. Apart from large labeled data, they require large processing time, robust hardware like GPUs. As the data on the web which is a primary source for these models, is updated continuously, it posses challenges. The financial and infrastructure cost associated with this process requires assessing the feasibility for general public.

5.3. Ethical Consideration

As described in the methodology section, this thesis relied on an existing external dataset for SA. The dataset was available to use openly at the time of writing this. As the data was real reviews of students, it was ensured that the input provided to models did not contain any sensitive information of the students. If any such information was found it was removed. The main ethical objective of the work was to utilize the data responsibly and devoid of unethical motivations. The prompts designed to assess model's understanding were structured to be unbiased and avoided any form of discrimination based on factors such as gender or other potentially sensitive attributes. Moreover, as the work involved use of an AI model, ethical considerations extended to its responsible use.

6. Conclusion

6.1. Summary Of Findings

The emergence of large language models (LLMs) and AI tools like ChatGPT has impacted multiple domains including education and has fundamentally transformed natural language processing (NLP) tasks, including sentiment analysis. The literature review of the SOTA methods for SA conducted in this work revealed that LLMs based on transformers are the most advanced methods. This thesis work comprehensively analyzed the competence of ChatGPT in the educational context by performing SA on reviews of students on courses and comparing it with the SOTA models that require fine-tuning on the dataset (BERT, XLNet) and those that don't (bart-large-mnli, RoBERTa-large-mnli) using quantitative metrics.

From the results it can be noted that in case of three labels, ChatGPT shows comparable performance to models that requires fine-tuning and better or similar performance to those that don't. In this case, XLNet and BERT show the highest accuracy of 95 % and 98 % in Train-test and SKF method respectively whereas ChatGPT shows decent performance with 92 % and 90 % respectively. In case of five labels, performance of all the models declined. However, fine-tune models conclusively show better results than ChatGPT while it shows moderate performance compared to latter models. Here, again XLNet and BERT illustrate the best results with the values of 80 % and 98 % for Train-test and SKF respectively whereas ChatGPT struggles with a difference of upto 34 % with BERT. However, ChatGPT does better than all models on unseen data which was not part of fine-tuning. This highlights its strength as an open domain model without the necessity of fine-tuning on the specific dataset which is an expensive process in terms of cost and resources.

With the help of 7 prompting techniques, this work also analyzed how well ChatGPT understands linguistic nuances in the given texts using qualitative approach. The observations suggest that out of the five linguistic nuances that were examined ChatGPT shows promising performance in terms of correctness, consistency, relevance and robustness on all barring Irony. Among the prompting techniques, CoT, with step by step instructions is the most effective method in the given context with the accuracy of 88 % and leading in all four metrics. ChatGPT demonstrates an impressive knowledge of the language and offers a benefit of being usable in any domain without specific training.

Overall, LLMs or tools based on them like ChatGPT has addressed the issues of the traditional methods in NLP and has surpassed them in performance. However, it brings its own challenges. prompt sensitivity, bias, randomness are some of the limitations of ChatGPT. As the educational landscape evolves, integration of AI appears imminent. By mitigating the challenges of ChatGPT and leveraging its strengths it is worth exploring further for insightful sentiment analysis which could help in improved and personalized learning.

6.2. Implications and Future Work

This thesis work shows preliminary potential of ChatGPT in SA. Using SA can improve the education process greatly. Teachers can assess how effective their teaching method is, whether and how the students understand the topics. Using the impressive knowledge of the language that ChatGPT has demonstrated, it can be used to analyze students reviews beyond the most widely used method of labeling the reviews. It can be asked to predict the exact sentiments in the texts. The results although show the need for improvements in the models, they also indicate that this area is worth exploring with AI. For this, solutions can be explored to integrate ChatGPT in the learning platforms and to monitor its performance.

This study also has implications in the research field. This work can be extended to future versions of ChatGPT with the later models. The three labels - Negative, Neutral, Positive or five labels - Very Negative, Negative, Neutral, Positive, Very Positive could be replaced by precise sentiments. Considering the limitations of this work, the future work can experiment with different languages as well as datasets in the educational context. Regarding prompt engineering, more linguistic nuances which are complex could be included in the testing, newer prompting techniques or different prompts with existing techniques might provide new results. For more predictable behavior and generalizable conclusions regarding ChatGPT it needs to be tested with larger datasets.

References

[1] S. Abadal, A. Jain, R. Guirado, J. López-Alonso, and E. Alarcón. Computing Graph Neural Networks: A survey from algorithms to accelerators. *ACM Computing Surveys (CSUR)*, 54(9):1–38, 2021.

[2] B. Abibullaev, A. Keutayeva, and A. Zollanvari. Deep learning in EEG-based BCIs: a comprehensive review of transformer models, advantages, challenges, and applications. *IEEE Access*, 2023.

[3] M. M. Amin and B. W. Schuller. On prompt sensitivity of ChatGPT in affective computing. *arXiv preprint arXiv:2403.14006*, 2024.

[4] R. Catelli, L. Bevilacqua, N. Mariniello, V. S. di Carlo, M. Magaldi, H. Fujita, G. De Pietro, and M. Esposito. Cross Lingual Transfer Learning for Sentiment Analysis of Italian TripAdvisor reviews. *Expert Systems with Applications*, 209:118246, 2022.

[5] J. Y.-L. Chan, K. T. Bea, S. M. H. Leow, S. W. Phoong, and W. K. Cheng. State of the art: a review of Sentiment Analysis based on Sequential Transfer Learning. *Artificial Intelligence Review*, 56(1):749–780, 2023.

[6] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, et al. A survey on evaluation of Large Language Models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45, 2024.

[7] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu. Boosting for Transfer Learning. In *Proceedings of the 24th international conference on Machine learning*, pages 193–200, 2007.

[8] N. Das, B. Sadhukhan, R. Chatterjee, and S. Chakrabarti. Integrating Sentiment Analysis with Graph Neural Networks for enhanced stock prediction: A comprehensive survey. *Decision Analytics Journal*, page 100417, 2024.

[9] R. Das and T. D. Singh. Multimodal Sentiment Analysis: a survey of methods, trends, and challenges. *ACM Computing Surveys*, 55(13s):1–38, 2023.

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[12] T. G. Dietterich. Ensemble methods in Machine Learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer, 2000.

[13] S. Ekin. Prompt engineering for ChatGPT: a quick guide to techniques, tips, and best practices. *Authorea Preprints*, 2023.

[14] F. Eyben, M. Wöllmer, and B. Schuller. Openear—introducing the Munich open-source emotion and affect recognition toolkit. In *2009 3rd international conference on affective computing and intelligent interaction and workshops*, pages 1–6. IEEE, 2009.

[15] A. Gandhi, K. Adhvaryu, and V. Khanduja. Multimodal Sentiment Analysis: review, application domains and future directions. In *2021 IEEE Pune section international conference (PuneCon)*, pages 1–5. IEEE, 2021.

[16] S. Gattupalli. The art and science of Promptgramming. 2024.

[17] A. Ghorbanali and M. K. Sohrabi. A comprehensive survey on Deep Learning-based approaches for Multimodal Aentiment Analysis. *Artificial Intelligence Review*, 56(Suppl 1):1479–1512, 2023.

[18] L. Giray. Prompt engineering with ChatGPT: a guide for academic writers. *Annals of biomedical engineering*, 51(12):2629–2633, 2023.

[19] A. Go, R. Bhayani, and L. Huang. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009, 2009.

[20] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, et al. Challenges in representation learning: A report on three Machine Learning contests. In *Neural information processing: 20th international conference, ICONIP 2013, daegu, korea, november 3-7, 2013. Proceedings, Part III 20*, pages 117–124. Springer, 2013.

[21] S. Hochreiter and J. Schmidhuber. Long Short-Term Memory. *Neural computation*, 9(8):1735–1780, 1997.

[22] J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.

[23] J. Jia, W. Liang, and Y. Liang. A review of hybrid and ensemble in Deep Learning for Natural Language Processing. *arXiv preprint arXiv:2312.05589*, 2023.

[24] S. Karen. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv: 1409.1556*, 2014.

[25] J. Kazmaier and J. H. Van Vuuren. The power of Ensemble Learning in Sentiment Analysis. *Expert Systems with Applications*, 187:115819, 2022.

[26] K. Kheiri and H. Karimi. Sentimentgpt: Exploiting GPT for advanced Sentiment Analysis and its departure from current Machine Learning. *arXiv preprint arXiv:2307.10234*, 2023.

[27] B. Khemani, S. Patil, K. Kotecha, and S. Tanwar. A review of Graph Neural Networks: concepts, architectures, techniques, challenges, datasets, applications, and future directions. *Journal of Big Data*, 11(1):18, 2024.

[28] R. Kora and A. Mohammed. An enhanced approach for Sentiment Analysis based on meta-ensemble Deep Learning. *Social Network Analysis and Mining*, 13(1):38, 2023.

[29] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

[30] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett. The computer expression recognition toolbox (CERT). In *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pages 298–305. IEEE, 2011.

[31] R. Liu, Y. Shi, C. Ji, and M. Jia. A survey of Sentiment Analysis based on Transfer Learning. *IEEE access*, 7:85401–85412, 2019.

[32] 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. *arXiv preprint arXiv:1907.11692*, 2019.

[33] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 ieee computer society conference on computer vision and pattern recognition-workshops*, pages 94–101. IEEE, 2010.

[34] M. J. Lyons, S. Akamatsu, M. Kamachi, J. Gyoba, and J. Budynek. The Japanese Female Facial Expression (JAFFE) database. In *Proceedings of third international conference on automatic face and gesture recognition*, pages 14–16, 1998.

[35] A. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. Learning word vectors for Sentiment Analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150, 2011.

[36] H. H. Mao. A survey on self-supervised pre-training for Sequential Transfer Learning in Neural Networks. *arXiv preprint arXiv:2007.00800*, 2020.

[37] W. Medhat, A. Hassan, and H. Korashy. Sentiment Analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113, 2014.

[38] G. Meena, K. K. Mohbey, A. Indian, and S. Kumar. Sentiment Analysis from images using VGG19 based Transfer Learning approach. *Procedia Computer Science*, 204:411–418, 2022.

[39] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[40] J. Oppenlaender, R. Linder, and J. Silvennoinen. Prompting AI art: An investigation into the creative skill of prompt engineering. *arXiv preprint arXiv:2303.13534*, 2023.

[41] S. J. Pan and Q. Yang. A survey on Transfer Learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[42] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[43] K. Pipalia, R. Bhadja, and M. Shukla. Comparative analysis of different transformer based architectures used in Sentiment Analysis. In *2020 9th international conference system modeling and advancement in research trends (SMART)*, pages 411–415. IEEE, 2020.

[44] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, and S. Manandhar. SemEval-2014 task 4: Aspect based sentiment analysis. In P. Nakov and T. Zesch, editors, *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland, Aug. 2014. Association for Computational Linguistics.

[45] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[46] S. Raschka. Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint arXiv:1811.12808*, 2018.

[47] P. P. Ray. Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, 2023.

[48] M. T. Ribeiro, S. Singh, and C. Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[49] S. Rosenthal, N. Farra, and P. Nakov. Semeval-2017 task 4: Sentiment Analysis in twitter. *arXiv preprint arXiv:1912.00741*, 2019.

[50] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.

[51] T. Susnjak. Applying BERT and ChatGPT for Sentiment Analysis of Lyme disease in scientific literature. In *Borrelia burgdorferi: Methods and Protocols*, pages 173–183. Springer, 2024.

[52] V. Taecharungroj. "what can ChatGPT do?" analyzing early reactions to the innovative AI chatbot on twitter. *Big Data and Cognitive Computing*, 7(1):35, 2023.

[53] K. L. Tan, C. P. Lee, K. M. Lim, and K. S. M. Anbananthen. Sentiment Analysis with Ensemble Hybrid Deep Learning model. *IEEE Access*, 10:103694–103704, 2022.

[54] A. Tlili, B. Shehata, M. A. Adarkwah, A. Bozkurt, D. T. Hickey, R. Huang, and B. Agyemang. What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart learning environments*, 10(1):15, 2023.

[55] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[56] J. Wang, E. Shi, S. Yu, Z. Wu, C. Ma, H. Dai, Q. Yang, Y. Kang, J. Wu, H. Hu, et al. Prompt Engineering for healthcare: Methodologies and applications. *arXiv preprint arXiv:2304.14670*, 2023.

[57] M. Wang and G. Hu. A novel method for twitter Sentiment Analysis based on Attentional-Graph Neural Network. *Information*, 11(2):92, 2020.

[58] Y. Wang and Z. Luo. Enhance multi-domain Sentiment Analysis of review texts through prompting strategies. In *2023 International Conference on High Performance Big Data and Intelligent Systems (HDIS)*, pages 1–7. IEEE, 2023.

[59] Z. Wang, Q. Xie, Y. Feng, Z. Ding, Z. Yang, and R. Xia. Is ChatGPT a good sentiment analyzer? a preliminary study. *arXiv preprint arXiv:2304.04339*, 2023.

[60] J. Wu, X.-Y. Chen, H. Zhang, L.-D. Xiong, H. Lei, and S.-H. Deng. Hyperparameter optimization for machine learning models based on Bayesian optimization. *Journal of Electronic Science and Technology*, 17(1):26–40, 2019.

[61] R. Xu, J. Xu, and X. Wang. Instance level Transfer Learning for Cross Lingual Opinion Analysis. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011)*, pages 182–188, 2011.

[62] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le. XLNet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems*, 32, 2019.

[63] A. K. Zarandi and S. Mirzaei. A survey of Aspect-Based Sentiment Analysis classification with a focus on Graph Neural Network methods. *Multimedia Tools and Applications*, pages 1–77, 2023.

[64] X. Zhai. ChatGPT user experience: Implications for education. *Available at SSRN 4312418*, 2022.

[65] S. Zhang, H. Tong, J. Xu, and R. Maciejewski. Graph Convolutional Networks: a comprehensive review. *Computational Social Networks*, 6(1):1–23, 2019.

[66] W. Zhang, Y. Deng, B. Liu, S. J. Pan, and L. Bing. Sentiment Analysis in the era of Large Language Models: A reality check. *arXiv preprint arXiv:2305.15005*, 2023.

[67] X. Zhang, J. Zhao, and Y. LeCun. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28, 2015.

[68] Z.-H. Zhou. *Ensemble Learning*, pages 411–416. Springer US, Boston, MA, 2015.

[69] L. Zhu, Z. Zhu, C. Zhang, Y. Xu, and X. Kong. Multimodal Sentiment Analysis based on fusion methods: A survey. *Information Fusion*, 95:306–325, 2023.

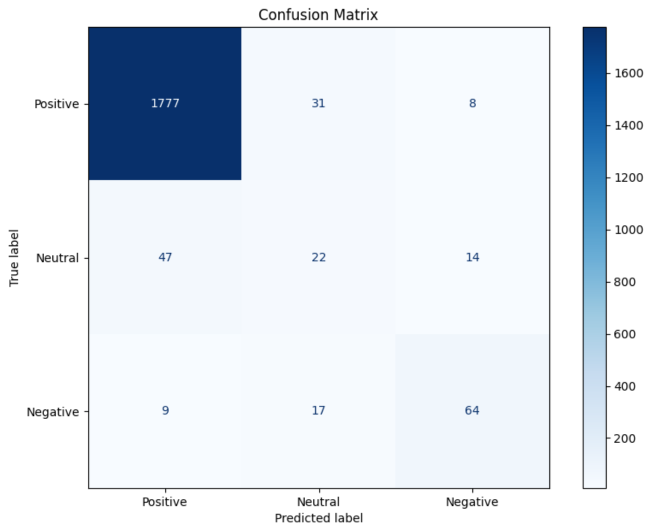
A. Appendix

A.1. Source Files

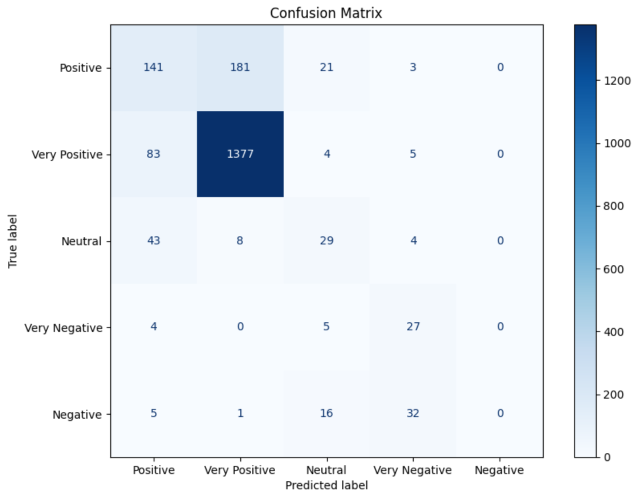
The files for code, dataset and results can be found in the following svn repository. Click **here** for the files.

A.2. Confusion Matrices

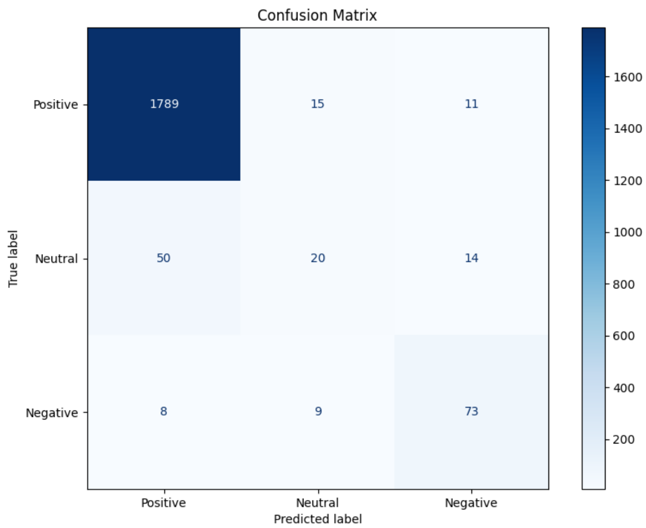
Following are the confusion matrices of predictions in SA using all the models. The numbers are absolute numbers which correspond to the actual number of samples.



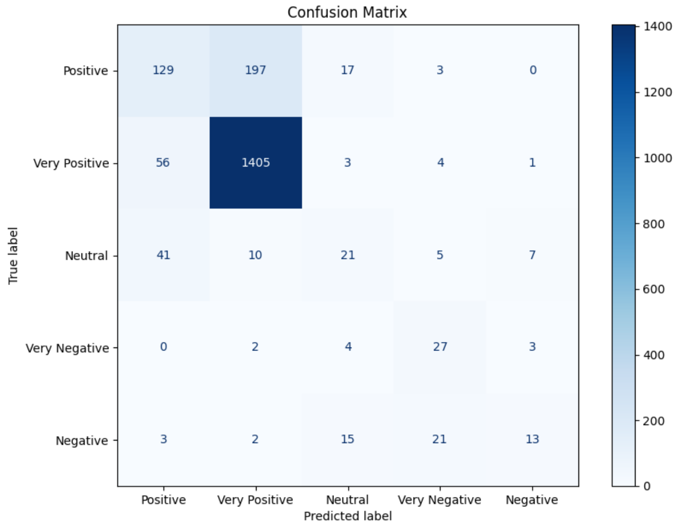
Confusion Matrix of Fine-tuned BERT using train-test in case of Three Labels



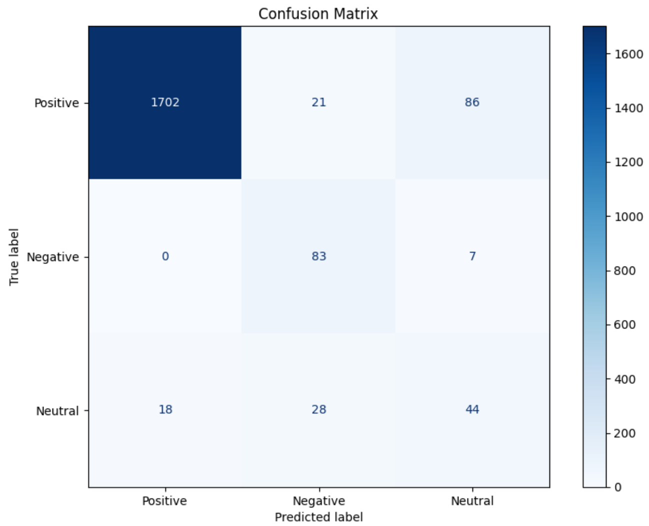
Confusion Matrix of Fine-tuned BERT using train-test in case of Five Labels



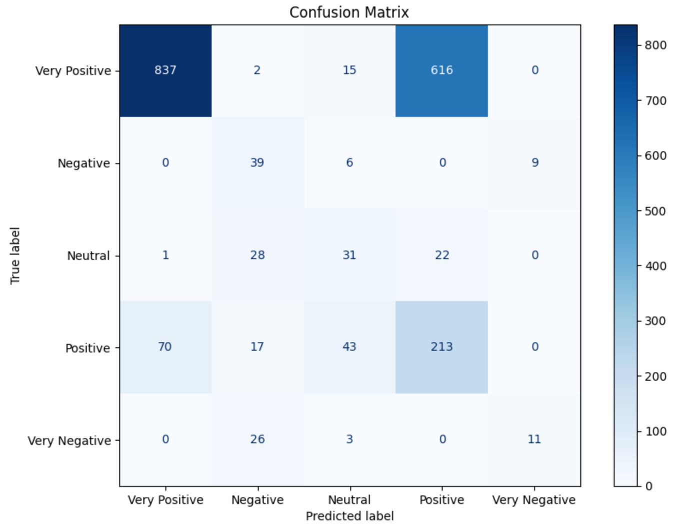
Confusion Matrix of Fine-tuned XLNet using train-test in case of Three Labels



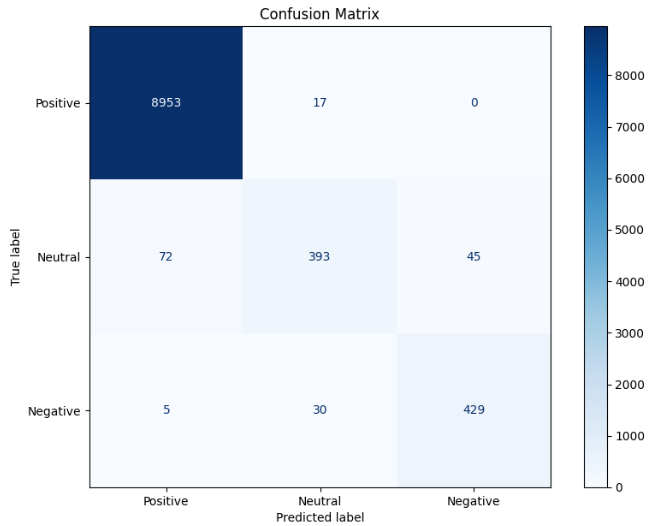
Confusion Matrix of Fine-tuned XLNet using train-test in case of Five Labels



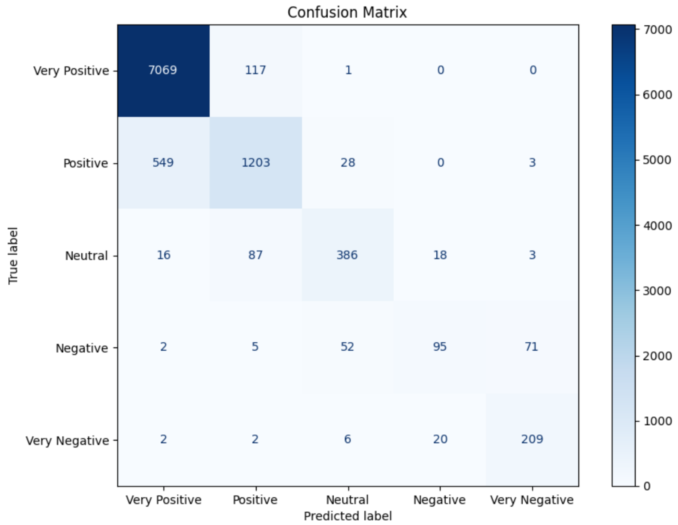
Confusion Matrix of ChatGPT when test part of the split was used in case of Three Labels



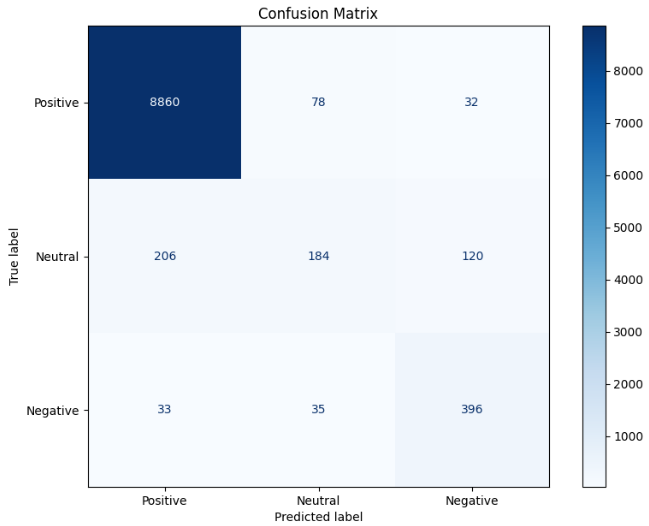
Confusion Matrix of ChatGPT when test part of the split was used in case of Five Labels



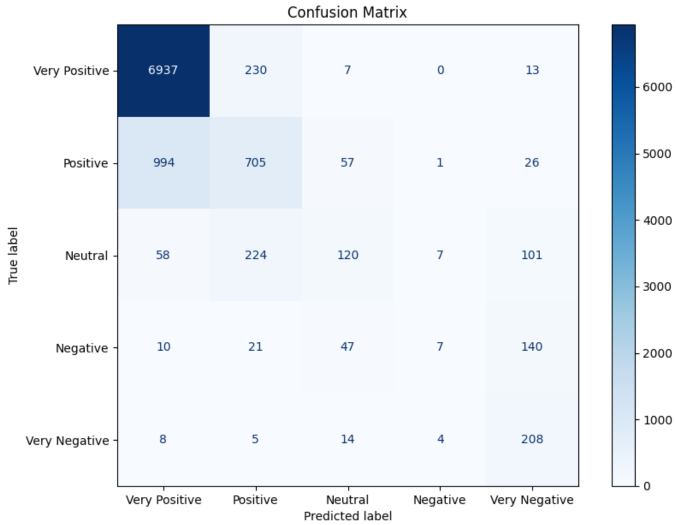
Confusion Matrix of Fine-tuned BERT using SKF in case of Three Labels



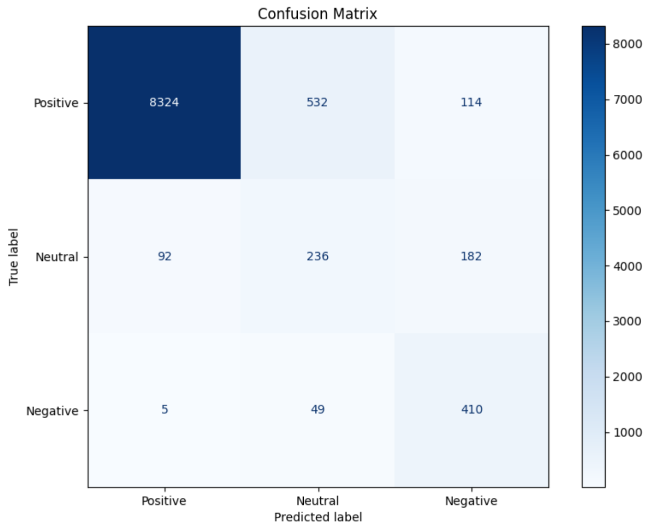
Confusion Matrix of Fine-tuned BERT using SKF in case of Five Labels



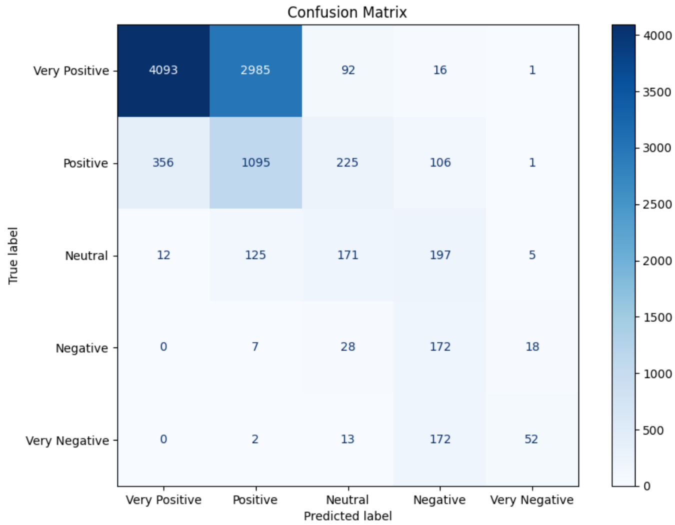
Confusion Matrix of Fine-tuned XLNet using SKF in case of Three Labels



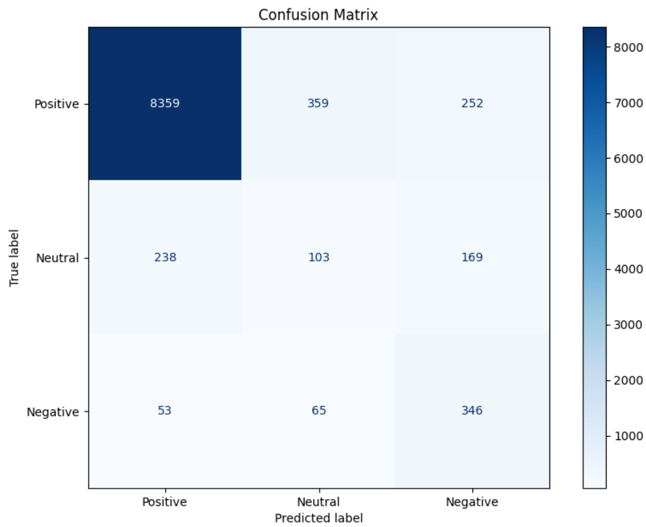
Confusion Matrix of Fine-tuned XLNet using SKF in case of Five Labels



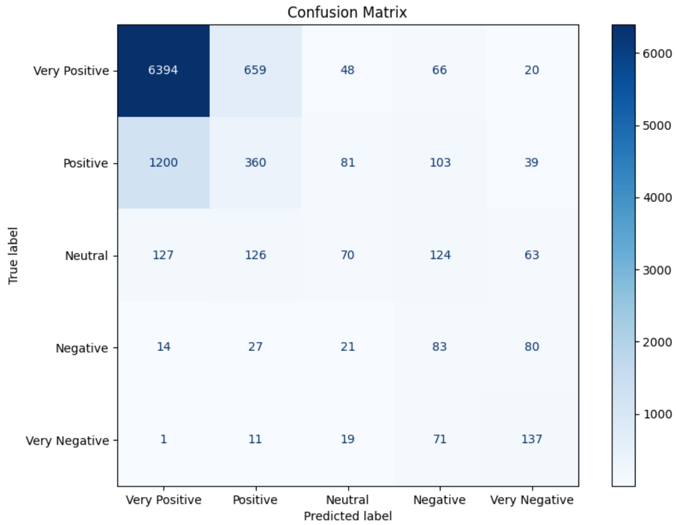
Confusion Matrix of ChatGPT when entire dataset was used in case of Three Labels



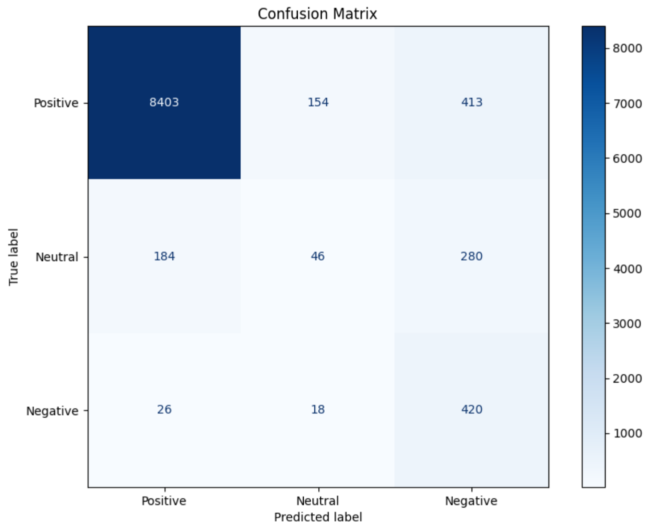
Confusion Matrix of ChatGPT when entire dataset was used in case of Five Labels



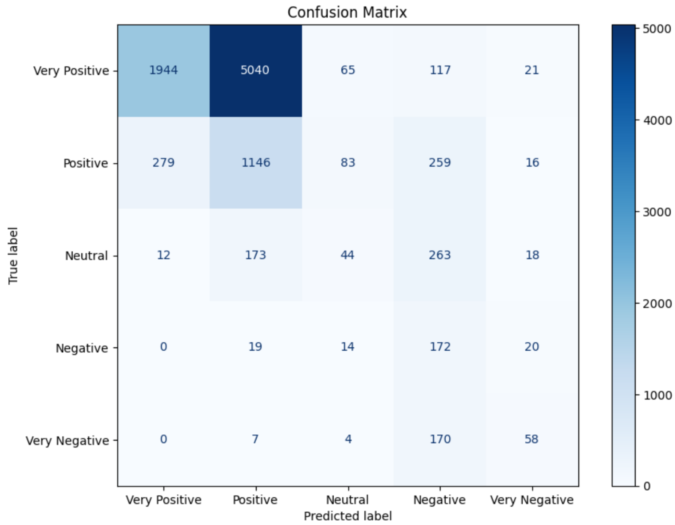
Confusion Matrix of bart-large-mnli in case of Three Labels



Confusion Matrix of bart-large-mnli in case of Five Labels



Confusion Matrix of RoBERTa-large-mnli in case of Three Labels



Confusion Matrix of RoBERTa-large-mnli in case of Five Labels