EAI Endorsed Transactions
on Scalable Information Systems
Research Article EAI.EU

Manifesto of Deep Learning Architecture for Aspect Level Sentiment Analysis to extract customer criticism

N. Kushwaha[1], B. Singh[1,*] and S. Agrawal[2]

[1]CSE Department, IIIT Ranchi, India
[2]Bennet University, India

Abstract

Sentiment analysis, a critical task in natural language processing, aims to automatically identify and classify the sentiment expressed in textual data. Aspect-level sentiment analysis focuses on determining sentiment at a more granular level, targeting specific aspects or features within a piece of text. In this paper, we explore various techniques for sentiment analysis, including traditional machine learning approaches and state-of-the-art deep learning models. Additionally, deep learning techniques has been utilized to identifying and extracting specific aspects from text, addressing aspect-level ambiguity, and capturing nuanced sentiments for each aspect. These datasets are valuable for conducting aspect-level sentiment analysis. In this article, we explore a language model based on pre-trained deep neural networks. This model can analyze sequences of text to classify sentiments as positive, negative, or neutral without explicit human labeling. To evaluate these models, data from Twitter's US airlines sentiment database was utilized. Experiments on this dataset reveal that the BERT, RoBERTA and DistilBERT model outperforms than the ML based model in accuracy and is more efficient in terms of training time. Notably, our findings showcase significant advancements over previous state-of-the-art methods that rely on supervised feature learning, bridging existing gaps in sentiment analysis methodologies. Our findings shed light on the advancements and challenges in sentiment analysis, offering insights for future research directions and practical applications in areas such as customer feedback analysis, social media monitoring, and opinion mining.

Keywords: Sentiment Analysis, Decision Making, GRU, Airline Data, Social Media, Deep Learning Architecture, Glove, Capsule Network, BERT

Received on 26 December 2023, accepted on 02 April 2024, published on 09 April 2024

doi: 10.4108/eetsis.5698

*Corresponding author. Email: bsingh@iiitranchi.ac.in

1. Introduction

Sentiment analysis is a kind of data analysis in which the polarity of comments is estimated. By doing machine learning or deep learning methods these polarities are uncovered. In the process of data analysis, its involves collection, cleaning, transformation and modelling of data to capture hidden polarities for different decision making. Figure 1 show the required data analysis process.

Data are collected from numerous social media platforms, including Facebook, WhatsApp, LinkedIn, Twitter, Google Plus, YouTube, and Instagram, have gained widespread popularity [1] [2] [3]. Millions of users actively engage with these platforms to share their opinions and perspectives. When individuals plan to book tickets, they often rely on the ratings and feedback available on social media sites like Twitter and Facebook to inform their decision-making process. Consequently, companies are interested in employing techniques or tools that can effectively analyze passenger feedback. One such technique is the sentiment analysis [4] [5] [6].

Sentiment analysis is a type of classification task. In this the block of text is tested to check whether it is positive, negative or neutral. The important goal will be to analyse crowd interest in such a way that in will help to understand business requirements as per the crowd interest. It can be considered as a contextual mining of comments that indicate the social sentiment of a product or item. Figure 2 shows the required sentiment analysis. Sentiment analysis is a very active area of

research in natural language processing that allows for the extraction of opinions from a set of documents. Sentiment analysis can be investigated at various levels [4] [7] [9] [21]. Different machine learning (ML) algorithms have been utilized to determine the most suitable algorithm for the specific problem] [10] [11] [21].

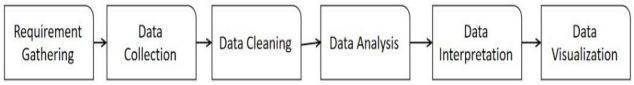


Figure 1. The abstract data analysis process

The performance evaluation involved analyzing the confusion matrix and accuracy of these algorithms. To gain valuable insight from a large number of reviews, the reviews must be categorized into positive and negative sentiment. Sentiment analysis, also known as opinion mining, is a natural language processing technique that involves determining the sentiment or emotional tone expressed in a piece of text [12] [13]. It aims to understand and classify the subjective opinions, attitudes, and emotions conveyed by individuals or groups towards a particular topic, product, service, or event. Sentiment analysis can be applied to various forms of text data, including social media posts, customer reviews, survey responses, and news articles. It helps businesses, organizations, and researchers gain insights into public opinion, customer feedback, and brand reputation, enabling them to make informed decisions, improve products or services, and tailor marketing strategies [9] [10].

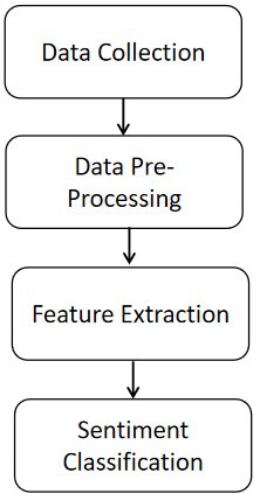


Figure 2. An abstract Sentiment Analysis process

Sentiment Analysis was used to categories over 9,000,00 reviews into positive and negative sentiments in the proposed work. For review classification, the Nave Bayes and Decision Tree (DT) classification models were used. Sentiment analysis has a wide range of applications, from determining customer attitudes towards products and services to determining voters' reactions to political advertisements [2] [14] [15]. Twitter is being widely used daily by people over the years to express views and sentiments. In airline industry, large number of customers post their views regarding services of the airlines like bag lost, good food, flight delay and many others. This helps airlines cater customers based on their reviews. In this paper we classify the dataset of review sentiments as Positive, Neutral, and Negative using ML techniques [4] [12] [9]. The structure of the paper is as follow: in section 2 literature review about sentiment analysis has given. The approach utilized to enhance the sentiment analysis, proposed framework and dataset details in section 3. In section 4 BERT model with pre-training. Result and analysis are given in section 5 and conclusion in section 6.

2. Literature Review

Sentiment analysis is a popular research topic in the field of natural language processing and has many applications in various industries. In this paper, four state of the arts classifiers, like DT, Logistic Regression (LR), Bayesian Naïve and Random Forest (RF), were used to compare the results of sentiment of text data over proposed BERT based sentiment analysis. In order to further enhance the accuracy and effectiveness of the sentiment analysis, it is important to explore the latest research and advancements in this area [16] [7] [17] [18] [19]. Furthermore, V. Hatzivassiloglou et al. [20] proposes a method for predicting the semantic orientation of adjectives using a corpus-based approach. The authors introduce a novel algorithm for identifying the semantic orientation of adjectives based on the co-occurrence patterns of words in the corpus. Qiu et al. [20] proposes a novel method for dissatisfaction-oriented advertising based on sentiment analysis. The authors use a ML approach to identify customer dissatisfaction and propose targeted advertising strategies to improve customer satisfaction.
Furthermore, S. Tan et al. [5] presents an empirical study of sentiment analysis for Chinese documents. The authors compare the performance of several ML algorithms for sentiment analysis, including Naïve Bayes, SVM, and DTs. Sentiment analysis has gained significant attention due to its wide range of applications. It is used in social media monitoring to understand public opinion and brand perception, customer feedback analysis to gauge user satisfaction, market research to track consumer sentiment, and many other domains. Various techniques are employed for sentiment analysis, including ML algorithms such as Naïve Bayes, LR, RF, and Support Vector Machines. Deep learning models, including recurrent neural networks (RNNs) and convolutional neural networks (CNNs), have also shown promising results in sentiment analysis tasks. The performance of sentiment analysis models is evaluated using metrics such as accuracy, precision, recall, and F1-score, among others. Researchers have explored feature engineering, sentiment lexicons, linguistic patterns, and domain adaptation techniques to enhance the accuracy and robustness of sentiment analysis models [22].

Yang et al [40] Investigates transformer-based models for language understanding, with potential applications in sentiment analysis. They have proposed the XLNet model for language understanding. DistillBERT, given by Sanh et al.,[41] explores model distillation techniques for reducing the size and computational cost of BERT-like models, which can impact sentiment analysis applications. In the paper [39] auther discusses visualization techniques for interpreting attention mechanisms in transformer models like BERT, providing insights into how sentiment information is processed.

S. Erevelles et al [1] discusses the use of big data and sentiment analysis in consumer analytics and marketing. The authors highlight the importance of sentiment analysis in understanding consumer preferences and behavior and propose a framework for using sentiment analysis in marketing strategies. S.Tong et al [16] presents a method for support vector machine active learning with applications to text classification. The authors propose a novel approach for selecting informative examples to label in order to improve the performance of the classifier. In literature a strong sentiment analysis has been done using ML models, but they are lack behind in the aspect level sentiment analysis that we have done through the BERT method. Here, we proposed an NLP model with multiple embedding techniques based on ML. A transformer-based bidirectional encoder representation (BERT) for extracting latent linguistic features from airline ratings. The detailed and complete information can be seen in the [42] [43] [44]. This study uses ML based approach along with BERT, RoBERTa and DistilBERT and information visualization techniques to investigate how feedback affects customer satisfaction in various aspects of flight service. The unrated aspects of airline reviews are then predicted from the rated aspects. In the next section a brief description of various analysis level which are available in literature has been provided.

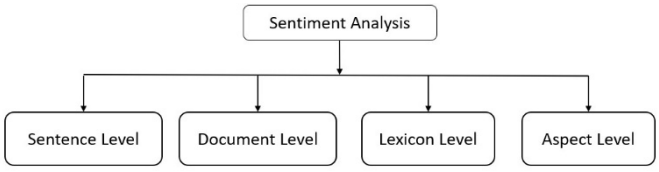


Figure 3. Various levels of sentiment analysis

2.1. Levels of Analysis

Sentiment analysis can be applied across three distinct levels: document, sentence, and aspect. Further details on each level will be expounded upon in the subsequent paragraphs. In Figure 3. categorization of sentiment analysis has been shown.

2.1.1 Document-level

Document-level analysis treats the entire text document as the primary unit of analysis [25]. This simplified approach assumes that the entire document reflects the opinion of a single entity. However, document analysis encounters challenges, particularly the presence of multiple and varied opinions within a document, sometimes conveyed through implicit language [26]. Often, documents undergo revisions at the sentence or aspect level before establishing the overall polarity of the entire text document.

2.2.2 Sentence-Level:

Sentence-level analysis focuses on individual sentences within a text, primarily applied in subjectivity classification. In text documents, sentences can be categorized into those expressing opinions and those that do not. Subjectivity classification involves assessing individual sentences to identify whether they convey facts or emotions and opinions. The primary objective of subjectivity classification is to filter out sentences devoid of sentiment or opinion [26].

2.2.3 Aspect-Level:

It is alternatively referred to as entity-level or feature-level analysis. Aspect-level analysis poses a significant challenge within the field of sentiment analysis. This approach involves evaluating sentiments related to specific entities and their respective aspects within a text document, rather than focusing solely on the overall sentiment of the entire document. Despite classifying the general sentiment of a document as positive or negative, the opinion holder may hold differing opinions about specific aspects of an entity [26]. To gauge opinions at the aspect level, it is imperative to identify the specific aspects of the entity. Valdivia et al. [27] emphasized the value of aspect-based sentiment analysis for business managers, as it allows for the extraction of customer opinions in a transparent manner. They also highlighted the ongoing challenge of detecting ironic expressions in TripAdvisor and advocated for a more comprehensive approach to review labeling beyond user ratings.

2.2.4 Lexicon-based

Lexicon-based learning represents a conventional method in sentiment analysis. This approach involves scanning documents for words that convey either positive or negative sentiments to humans. A predefined lexicon defines these words, eliminating the need for learning data in this method [42].

2.2.5 Hybrid models

In the realm of sentiment classification, hybrid models amalgamate the lexicon-based approach with machine learning techniques [24] [40] to formulate a lexicon-enhanced classifier. Lexicons play a crucial role in delineating domain-related features that serve as input for a machine learning classifier.

2.2. Research questions

The following research questions have been defined for this study:

- What features, both in terms of input and output, have been embraced in sentiment analysis?
- What approaches have been employed in sentiment analysis?
- Which domains have been covered in the utilized datasets?
- What difficulties and unresolved issues exist in the context of sentiment analysis?

Materials and Method

In this section, we discuss the techniques for our proposed framework. First of all, in Figure 8 a framework has been shown which represents the adopted methodology. Feature extraction and embedding method were done on training and testing data. TF-IDF is a scoring measures to reflect how relevant a term in the given document. For the embedding purpose Glove has been utilized which encode the co-occurrence probability ration between two worlds.

Data Set Description

The datasets used in this paper is taken from social media platform. Comments data that are included in this work are about six airlines i.e. Unites State, Delta, US Airways, United, Southwest and Vergin America [8]. Passenger ratings are recorded and categorized as positive, negative or neutral. Negative reviews are defined based on things like bad flights, flight delays, customer service issues, damaged luggage, flight cancellations or booking issues [8].
Positive ratings are defined based on fast flights, great flights, great flights, good brands, etc. The descriptive analysis has been carried out that we have shown in Figure 4, Figure 5, 6. Figure 4 shows the comments of customers reviews percentage as a pie chart. Figure 5 describes top ten negative reviews given by customers . It can be seen that most of customers faces service related problem of airlines.
Dataset used in this research is not a balanced data set that can be well understood from Figure 6a-b. It has a smaller number of positive comments in comparison to negative comments. The attributes of this datasets are tweet_id, airline_sentiment, Airline sentiment_confidence, airline, airline sentiment gold, name, retweet_count, location etc. In order to prepare the dataset for analysis, data pre-processing techniques were applied. This step is essential in ML to address potential issues arising from the nature of the dataset collected from social sites. Such data can be prone to inaccuracies and may lack certain attributes necessary for analysis. Thus, it is crucial to resolve these issues prior to conducting any further analysis.

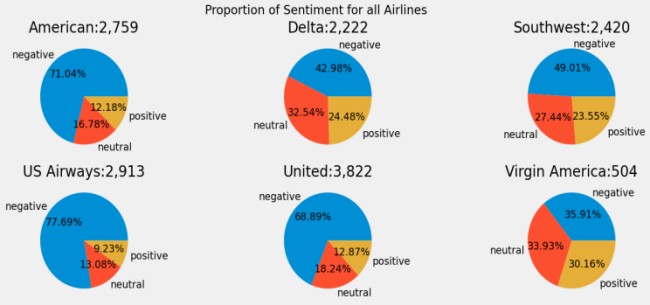


Figure 4. A pie chart showing the proportion of sentiments of all six airline companies.

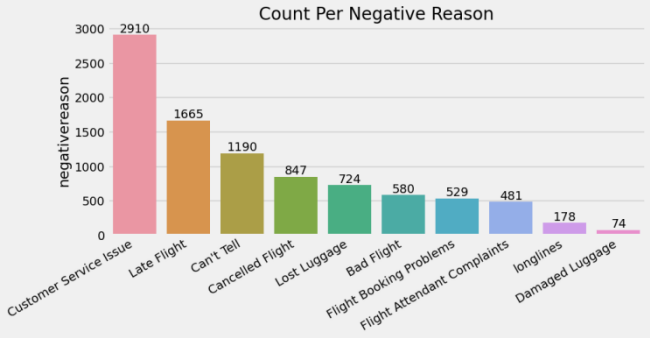
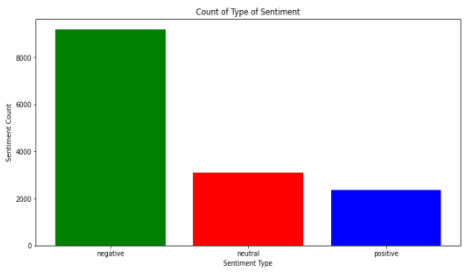
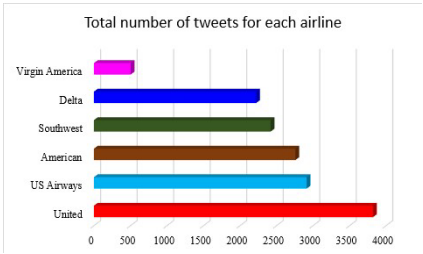


Figure 5. Count per top ten negative reasons





(a) **(b)**

Figure 6. a) Graph showing number of negative, positive and neutral comments/review in the data sets. (b) Bar Graph representing the number of reviews for each airline, in the x-axis it is number of reviews and y-axis represent the name of airlines.

In pre-processing some required columns are selected and some common text processing algorithms are performed to: Remove empty reviews, convert all the reviews to lower case, remove numbers, tweet account names, website URLs, special characters and white spaces. Figure 7 depicts the mood of passengers toward each airline companies. We observe that United, US Airways, American substantially get negative reactions and tweets for Virgin America are the most balanced. We have used two word embedding techniques namely, Word2seq and Glove that are described below.

Word2seq

A method based on word sequencing [32] was employed, neglecting the consideration of individual word weights. This approach involved transforming the word sequence into a matrix, where the length represented the input size, and the height corresponded to the number of observations.

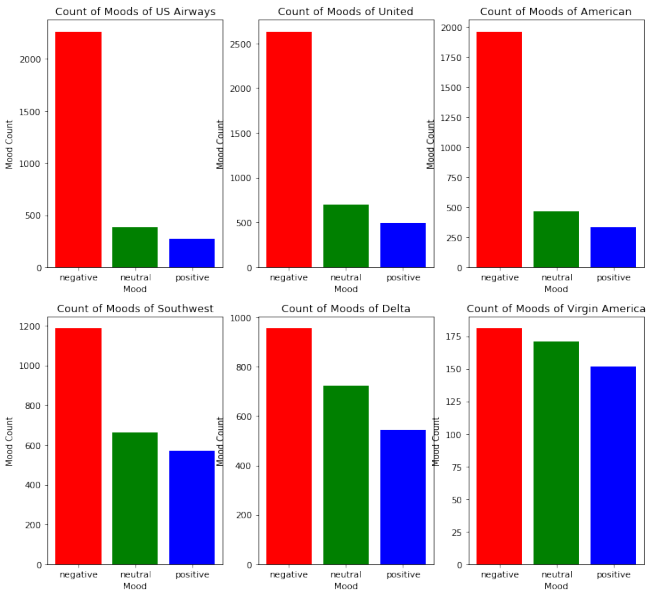


Figure 7. Count of mood as positive, negative and neutral of all six airlines. Virgin America is getting balanced feedback however rest of airline companies getting substantially negative reaction.

Glove

GloVe [33, 34] is an unsupervised learning algorithm that operates by projecting words into a meaningful space, facilitating the generation of vector representations for words. It is a technique used to represent words as dense vectors in a continuous vector space. It is based on the idea that words that frequently co-occur in similar contexts have similar meanings. GloVe constructs a co-occurrence matrix from a large corpus of text, where each element represents the frequency of occurrence of a pair of words within a specific context window. In this space, semantic similarity is intricately connected to the distance between words. In this work we have utilized these co-occurrence statistics to train a neural network to learn word embeddings that capture semantic relationships between words.

Methodology

In our research, we employed Machine Learning techniques like Naive Bayes, , Decision Tree, Logistic Regression, random forest and BERT that was considered in the conference. Now we have extended the analysis with BERT and its variations DistilBERT, and RoBERTa along with the all previous methods for sentiment analysis of airline customer feedback. We aimed to identify the most effective model by evaluating their performance using metrics such as accuracy, precision, recall, and the F-1 score. The complete procedure is illustrated in Figure 8. Initially, we utilized a tokenizer specific to each model to preprocess the text input. Next, the encoded text was converted into a tensor dataset, serving as input for the classification model. Here, we have used two encoding method glove and word2seq. The logits produced by the classification model were then translated into classified labels, which were ultimately evaluated based on metric performance. Complete outline of the proposed work has been shown in Figure 9.

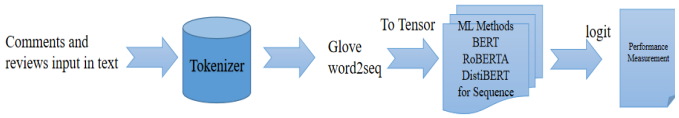


Figure. 8. Methodology that followed in this study

Evaluation Metrics

We present the evaluation metrics used in our work. For the performance evaluation, we utilized widely accepted metrics for example Precision, Recall. F1-score, Sensitivity, Specificity and Accuracy as given by equations from (1) – (5).

$$P=TP/(TP+FP) \quad (1)$$
$$R=TP/(TP+FN) \quad (2)$$
$$F1\text{-}Score=2*P*R/(P+R) \quad (3)$$
$$S=TN/(TN+FP) \quad (4)$$
$$Acc=TP+TN/(TP+FP+FN+TN) \quad (5)$$

Machine Learning Algorithms

We discuss four tradition ML methods that we have used in our study. Namely DT, LR, Naïve Bayes and RF. Here we are going for the briefing of these algorithm as these are the very well standard methods. As our motive was to analyze the aspect level sentiment analysis through ML algorithm. The analysis of results has been given in next section. RF [10] has demonstrated notable success in sentiment analysis tasks, outperforming various alternative ML methods. Its ability to handle high-dimensional data, manage noise, and capture complex relationships between features contributes to its effectiveness.

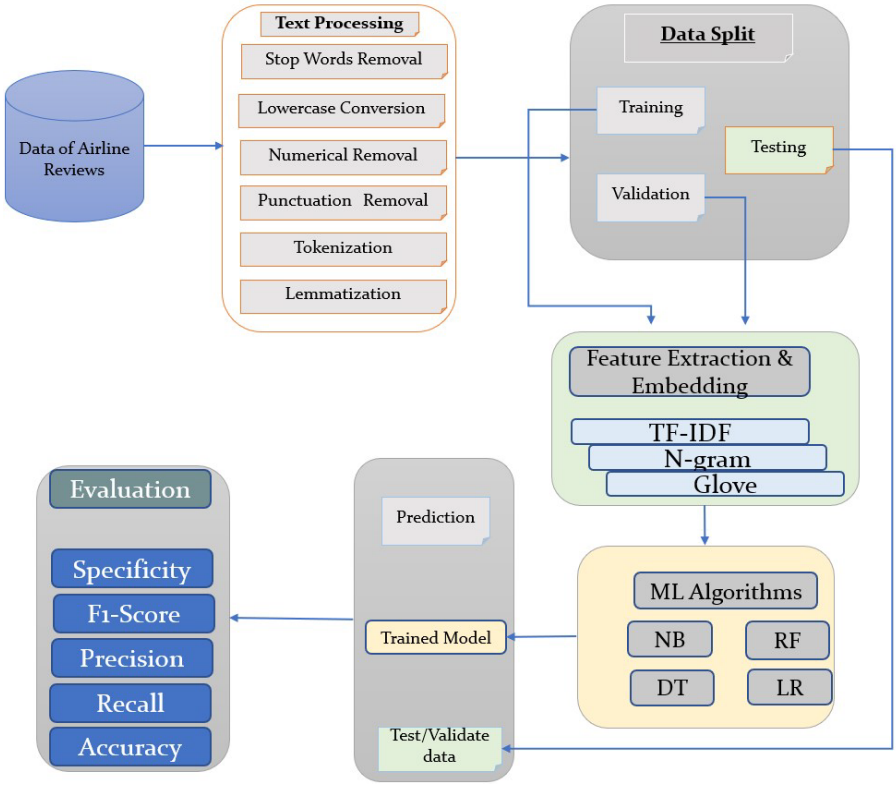


Figure 9. The complete outline of our proposed framework

Furthermore, its scalability and efficiency make it an attractive option for large-scale sentiment analysis applications. DT [19] have proven to be effective and interpretable models for sentiment analysis tasks. Their ability to handle both categorical and textual features, provide insights into feature importance, and offer robust performance makes them valuable in various application domains. However, challenges such as handling imbalanced data and adapting to evolving language patterns require further exploration and refinement.

Naïve Bayes, a probabilistic ML algorithm, has gained popularity due to its simplicity, efficiency, and competitive performance in sentiment analysis tasks. Its simplicity, competitive performance, and scalability make it a popular choice in various application domains. However, careful consideration of the feature independence assumption and its limitations in capturing complex relationships is essential for obtaining accurate sentiment analysis results [11]. LR, a widely-used statistical modeling technique, has shown promising results in sentiment analysis tasks. LR offers a well-established and interpretable approach for sentiment analysis tasks. Its ability to handle both binary and multiclass classification problems, along with its competitive performance in various application domains, makes it a valuable tool. However, its limited ability to capture complex nonlinear relationships and sensitivity to outliers should be considered when applying LR to sentiment analysis [9].

Deep Learning Algorithms

Deep learning models consist of intricate architectures comprising multiple layers of neural networks, systematically extracting high-level features from input data. Convolutional Neural Networks (CNN) utilize convolutional filters to detect patterns, widely applied in image recognition and, to a lesser extent, in Natural Language Processing (NLP). Recurrent Neural Networks (RNN) are crafted to recognize sequential patterns, demonstrating notable effectiveness in contexts where context is pivotal, making them particularly promising for sentiment analysis. Long Short-Term Memory (LSTM) networks, a specialized variant of RNN, excel in capturing long-term context and dependencies, proving especially beneficial in NLP tasks where extended dependencies are crucial. These aforementioned deep learning algorithms are acknowledged as promising techniques capable of elevating the performance of NLP tasks [23]. In our study, following methods of deep learning has been utilized to gain the polarities of data. The schematic architecture of Graph Neural Network, Memory Network, GRU and capsule network has been shown in Figure 10.

Graph Neural Network (GNNs)

Graph Neural Networks (GNNs) have gained attention for their ability to model complex relationships and dependencies in structured data. In the context of sentiment analysis, where relationships between words, phrases, and entities play a crucial role, GNNs offer a promising approach. Through GNNs, a graph representation, node embedding, message passing, and graph attention network can be generated.

Memory Network

Memory networks, also known as MemNets or Memory Augmented Networks, are a class of neural network architectures designed to incorporate memory and attention mechanisms. These networks are particularly effective in tasks that require reasoning over sequential or structured data [28]. In a Memory Network, an external memory matrix is utilized to store information from input sequences, and attention mechanisms are employed to selectively read from and write to this memory. This architecture enables the network to access and update information dynamically, making it well-suited for tasks involving long-term dependencies and complex reasoning. Memory networks, also known as MemNets or Memory Augmented Networks, have been widely discussed and expanded upon in the literature. Researchers have explored various aspects of memory networks, including their architectures, applications, and enhancements.

In this seminal paper by Weston et al [29], the authors introduced the concept of Memory Networks, outlining an architecture that utilizes an external memory matrix for storing and accessing information dynamically. The key innovation lies in the attention mechanisms that enable the network to selectively read from and write to the external memory, making Memory Networks well-suited for tasks involving sequential or structured data, where long-term dependencies and complex reasoning are essential. This method has ability for contextual understanding, handling long term dependencies, attention mechanism and learning representations. However, in this work it is explored as entity and aspect level sentiment analysis. It's important to note that the application of memory networks in sentiment analysis may vary based on the specific architecture used and the nature of the sentiment analysis task. Researchers continue to explore and refine memory-augmented models to enhance their effectiveness in handling sentiment-related challenges.

GRU (Gated Recursive Unit)

GRUs, considered as variants of LSTMs, function as LSTMs without an output gate, featuring two crucial gates: the update gate and the reset gate. GRUs are particularly noteworthy for their gating mechanisms, which enable them to capture and retain relevant information over sequential data. The architectural representation of a GRU is depicted in Fig. 10, and its gating mechanism can be elucidated as follows: Update Gate: This gate dictates the extent to which information (memory) should be retained for future use. The input, ht−1, signifies information from the preceding time step (t − 1 state), xt represents the current state value, and zt denotes the update gate value. Reset Gate: The reset gate, rt, determines how much historical information the network should discard. Calculate the Candidate Value: The candidate value, ht, is computed by taking the element-wise product (denoted by) of the reset gate rt and ht−1. This process determines the information to be omitted from previous time steps. It involves adding U xt, where U is a parameter vector, and applying the hyperbolic tangent (tanh) activation function to the output. Calculate the Final Memory Value: To obtain the ultimate memory value for the current unit, ht, update gate values z_t are essential. These values determine which information is required from ht and which information is needed from the previous value h_{t-1}.

Gated Recurrent Units (GRUs)applied effectively to sentiment analysis tasks, where the goal is to determine the sentiment expressed in a piece of text, such as positive, negative, or neutral [35].

Capsule Network

Geoffrey et al. [30] proposed Capsule Networks to address some limitations of traditional convolutional neural networks (CNNs), particularly in handling hierarchical relationships among features. The motivation behind Capsule Networks is to overcome the limitations of pooling layers in CNNs, which can lead to loss of spatial relationships and hierarchical information. Capsules, through dynamic routing, aim to capture the hierarchical structure of features in a more effective way. While Capsule Networks show promise in capturing hierarchical relationships, their effectiveness in sentiment analysis may depend on the specific dataset, task complexity, and the scale of training data. Using Capsule Networks (CapsNets) for sentiment analysis involves leveraging their ability to capture hierarchical relationships and spatial hierarchies in data. Capsule Networks can be adapted for aspect-based sentiment analysis, where the goal is to understand sentiment related to specific aspects or entities within a text. The hierarchical nature of CapsNets makes them suitable for capturing sentiment nuances associated with different aspects. Capsule Networks [31] need to be trained on labeled sentiment data to learn to effectively capture sentiment-related features. This involves adjusting the network's parameters to align with the sentiment labels provided in the training data.

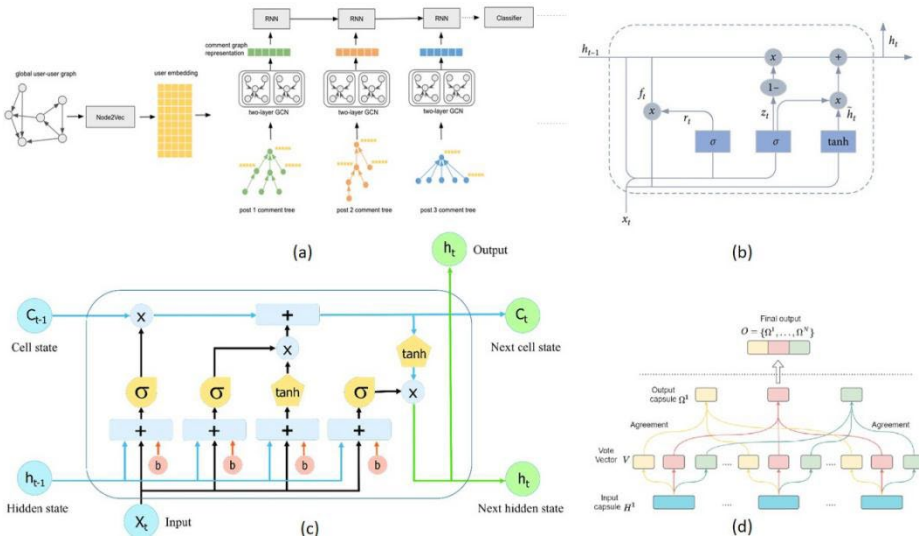


Figure 10. The schematic architecture of Graph Neural Network, Memory Network, GRU and capsule network.

BERT Model for Sentiment Analysis

BERT is a ML method based on transformers that Google developed for pre-training natural language processing (NLP). The Transformer language model, which has layers of self-aware heads and a variable number of encoders, is at the heart of BERT. The attention mechanism known as a Transformer, which is used by BERT, learns the contextual connections between words (or subwords) in text. Vanilla-style Transformers contain two separate mechanisms: an encoder that reads the text input and a decoder that creates predictions for the task [7] [13]. Since the purpose of BERT is to generate language models, we only need the Transformer's encoder mechanism. There are two variations of the pretrained BERT model. Both his BERT model sizes feature numerous encoder layers (referred to as transformer blocks in publications). 12 for the base version and 24 for the large version. as shown in Figure 11. Also, the pre-training model of BERT has given in Figure 12. BERT BASE and BERT LARGE refer to two different variations of the BERT model based on their model size and capacity.

Utilizing the Transformer, an attention mechanism that captures contextual relationships among words (or sub-words) in text, BERT, or Bidirectional Encoder Representations from Transformers, is designed. The Transformer comprises two distinct components—an encoder for processing text input and a decoder for task prediction. However, given BERT's objective of creating a language model, only the encoder mechanism is employed. BERT operates as a bidirectional transformer, pre-training extensively on vast amounts of unlabeled textual data to acquire a language representation applicable for fine-tuning in specific machine learning tasks. Despite its noteworthy performance surpassing the NLP state-of-the-art in various challenging tasks, BERT's success can be attributed to the bidirectional transformer, novel pre-training tasks like Masked Language Model and Next Structure Prediction, extensive data, and the computational power afforded by Google [37].

BERT BASE has 12 transformer layers, 12 attention heads, and a hidden size of 768, resulting in a total of approximately 110 million parameters. On the other hand, BERT LARGE has 24 transformer layers, 16 attention heads, and a hidden size of 1024, leading to around 340 million parameters. The larger model size of BERT LARGE allows it to capture more complex patterns and dependencies in the input data. During fine-tuning, BERT is further trained on specific downstream tasks with labeled data. This fine-tuning process adapts the pre-trained BERT model to perform task-specific operations, such as sentiment analysis, by adding task-specific layers on top of the BERT model. The fine-tuning stage allows the model to learn task-specific patterns and improve its performance on the target task. One key advantage of BERT is its ability to capture contextual information, which helps in understanding the meaning and nuances of words in different contexts. This contextualized representation is valuable for various NLP tasks, including sentiment analysis, as it allows the model to consider the surrounding words and sentences when making predictions. BERT BASE and BERT LARGE are pre-trained language models that leverage transformer-based architectures and self-attention mechanisms to capture contextual information. These models have been successfully applied to various NLP tasks, and their performance can be further enhanced through fine-tuning on specific downstream tasks.

The attention mechanism, a fundamental component ensuring the robust capability of transformers, was introduced by Vaswani [38] to address the challenge of handling long sequences in recurrent neural networks (RNNs). This mechanism computes attention scores between each element in the input sequence and the current element. Subsequently, the scores pass through a Softmax layer, resulting in attention weights. These weights are then

utilized to compute a weighted sum, generating a final context vector. This process enables transformers to effectively capture dependencies, both short-range and long-range, within extensive textual corpora. Equation (1) details the calculation of the attention score.

$$\text{Attention}\left(\text{Q},\text{K},\text{V}\right)=soft\max\left(\frac{QK^{T}}{\sqrt{d_k}}\right) \quad (7)$$

Here K represent the input key matrix, V is the value matrix, Q is the query matrix and d_k is the dimentionality of the keys.

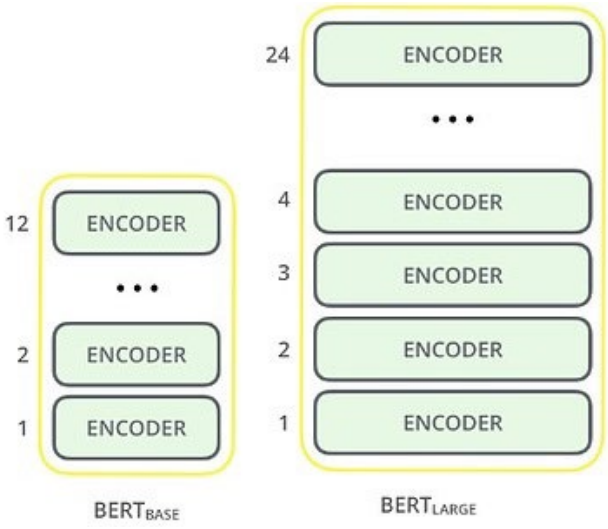


Figure 11. Two variant of BERT, BERTBASE and BERTLARGE with 12 and 24 number of encoders respectively

Generally, in text classification, Softmax layers deals the pivotal role in determining the likelihood of a data observation belonging to a specific class. This is achieved by inputting the first token of the sequence's final hidden state into the model. When BERT is employed for downstream tasks, it has the capacity to autonomously adjust its weights and adapt the output layer to suit the specific requirements of the task at hand. For instance, a Softmax layer is employed for tasks involving multi-label classification, whereas a sigmoid layer is employed for binary classification.

The mathematical expressions for the Softmax and sigmoid functions are presented in Equations (8) and (9). for vector Z=[Z1, Z2,……Zk] of k raw scores.

$$Soft\max:(z_t)=\frac{e^{z_i}}{\sum_{j=1}^{k}e^{z_j}} \quad \text{for i=1…..k} \quad (8)$$

$$Sigmoid:\sigma(Z)=\frac{1}{1+e^{-Z}} \quad (9)$$

RoBERTa

RoBERTa, means Robustly optimzed BERT approach, is a transformer-based pre-trained language model that has demonstrated exceptional performance across various natural language processing (NLP) tasks. RoBERTa employs dynamic masking during training, which involves masking different sets of words in each iteration. This dynamic masking strategy contributes to better contextualized representations [36]. The optimization of training strategies, such as the removal of NSP and the introduction of dynamic masking, has been instrumental in achieving superior performance compared to previous language models [36]. RoBERTa's architecture, rooted in the transformer model and enhanced by specific modifications such as dynamic masking and the removal of NSP, contributes to its robust performance in understanding contextual information within language data. Similar to BERT, RoBERTa utilizes masked language modeling as a pre-training objective. During pre-training, a percentage of input tokens are randomly selected and masked, and the model is trained to predict the masked tokens based on the context provided by the surrounding tokens. One notable modification in RoBERTa is the removal of the next sentence prediction (NSP) objective, which was part of BERT. By excluding NSP during pre-training, RoBERTa aims to improve model performance and encourage better understanding of context [36].
The RoBERTa approach suggested in this study comprises a pre-trained RoBERTa transformer followed by a bidirectional LSTM layer (BiLSTM). The representation vector, formed by concatenating the RoBERTa and LSTM outputs, undergoes pooling and is subsequently processed through a fully connected layer activated by softmax that is the target as given in the Figure 13.

DistilBERT

Victor Sanh et al. [45] proposed the DistilBERT. It is a smaller, distilled version of the BERT (Bidirectional Encoder Representations from Transformers) model. It retains much of the architecture and functionality of BERT but is smaller and faster, making it more suitable for deployment in resource-constrained environments or for applications where inference speed is critical. It has fewer layers and fewer attention heads compared to BERT, resulting in a smaller model size [41]. To harness the learned inductive biases from larger models during pre-training, the authors propose a triple loss mechanism that integrates language modeling, distillation, and cosine-distance losses.

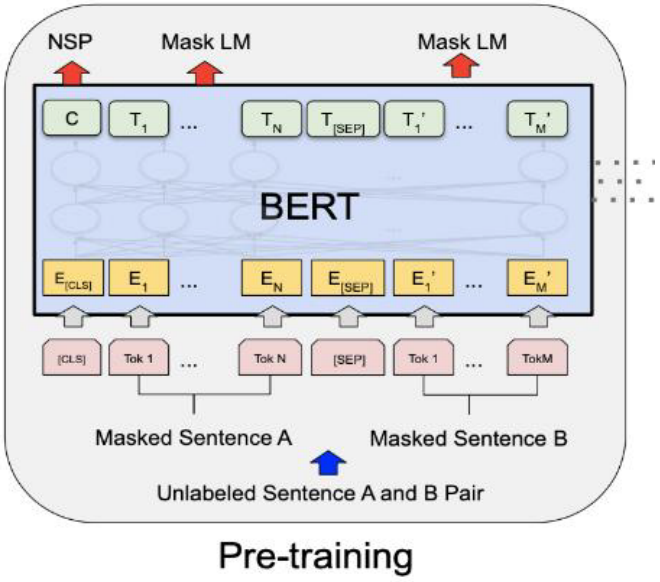


Figure 12. The diagram of Pre-training model of BERT

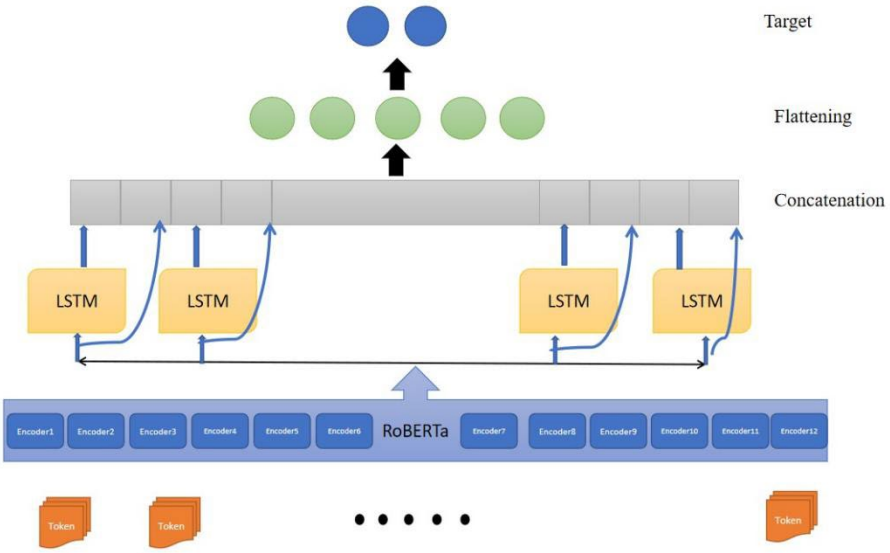


Figure 13. The diagram of Pre-training model of RoBERTa

```
Model: "model"
_________________________________________________________________________________________________
 Layer (type)                Output Shape              Param #   Connected to
=================================================================================================
 input_1 (InputLayer)        [(None, 128)]             0         []

 input_2 (InputLayer)        [(None, 128)]             0         []

 tf_bert_model (TFBertModel  TFBaseModelOutputWithPooli  109482240  ['input_1[0][0]',
 )                           ngAndCrossAttentions(last_            'input_2[0][0]']
                             hidden_state=(None, 128, 7
                             68),
                              pooler_output=(None, 768)
                             , past_key_values=None, hi
                             dden_states=None, attentio
                             ns=None, cross_attentions=
                             None)

 dense (Dense)               (None, 3)                 2307      ['tf_bert_model[0][1]']

=================================================================================================
Total params: 109484547 (417.65 MB)
Trainable params: 109484547 (417.65 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________________________________________
```

Figure. 14. Model summary of BERT

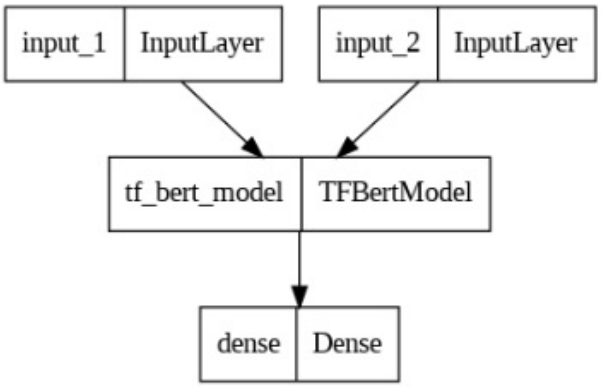


Figure. 15. Block architecture of BERT model

```
_________________________________________________________________________________________________
 Layer (type)                Output Shape              Param #   Connected to
=================================================================================================
 input_1 (InputLayer)        [(None, 80)]              0         []

 input_2 (InputLayer)        [(None, 80)]              0         []

 tf_roberta_model (TFRobert  TFBaseModelOutputWithPooli  124645632  ['input_1[0][0]',
 aModel)                     ngAndCrossAttentions(last_            'input_2[0][0]']
                             hidden_state=(None, 80, 76
                             8),
                              pooler_output=(None, 768)
                             , past_key_values=None, hi
                             dden_states=None, attentio
                             ns=None, cross_attentions=
                             None)

 dense (Dense)               (None, 3)                 2307      ['tf_roberta_model[0][1]']

=================================================================================================
Total params: 124647939 (475.49 MB)
Trainable params: 124647939 (475.49 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________________________________________
```

Figure. 16. Model summary of RoBERTa

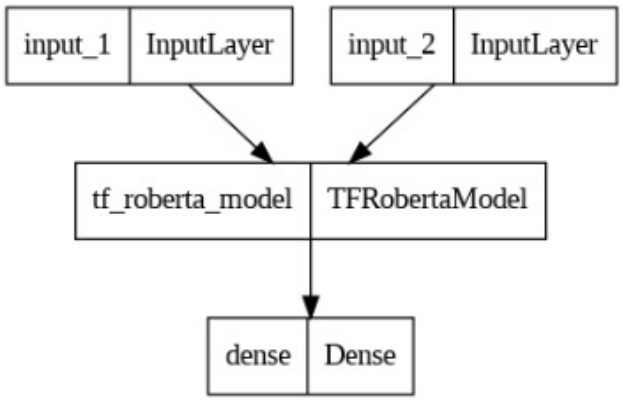


Figure. 17. Block architecture of RoBERTa model

Results and Discussion

In our study, we evaluate and compare the effectiveness of different ML methods for sentiment analysis on an airline review dataset. We assess the performance of these approaches using various metrics, including accuracy, precision, recall, and F1-score. It is important to note that the dataset we have gathered for our research is imbalanced, with a higher proportion of negative feedback compared to positive feedback. The comparison of all the ML models is shown in Table 1. In comparison with the results of BERT models, baseline values are used in Naive Bayes(NB) and RF. In the extended work, BERT, RoBERTa and DistiBERT has been tested on review data sets. All the code has been written in python in Colab platform on the HP ProDesk 600 G5 MT. Model summary and block diagram of BERT and RoBERTa model have been given in Figure 14-17.

Comparison of state-of-the-art-methods

We perform the statistical analysis of performance metrics. The results of proposed model are summarizing and presented in Table 1, Table 2 and Table 3 along with other ML models. Our estimations are based on the well k know measurement parameters like precision, recall, F1-score, sensitivity and accuracy. Table 1, showing the comparison of precision, recall and F1-score while in Table 2 we are depicting the accuracy, sensitivity and Specificity of four ML models. Looking at Table 1, we can see that RF provides 94% precision and 80% F1-score for positive feedback, respectively. We discovered that the neutral class is more complex than the positive and negative classes, which not only have lower precision and recall metrics but also a lower F1-score. Looking at the BERT model's performance, we see that it has an accuracy of 94%, with the 92% F1-score on the positive class and the lowest F1-score on the neutral class.

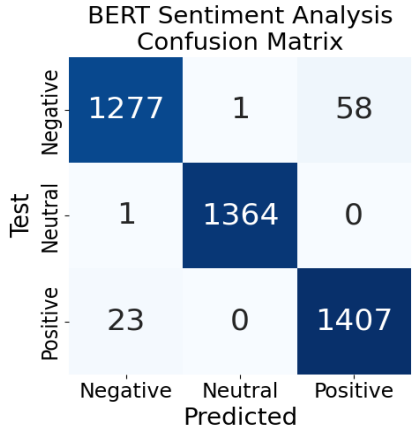


Figure. 18. Confusion matrix of BERT model

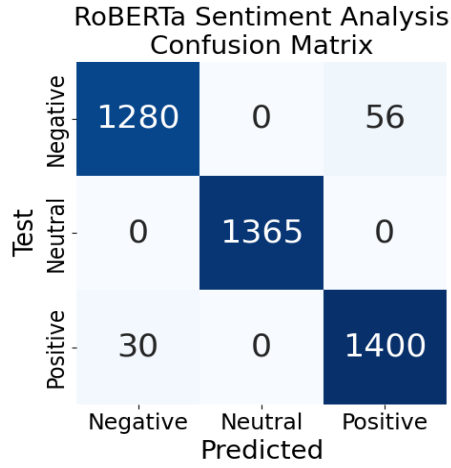


Figure. 19. Confusion matrix of RoBERTa model

RoBERTa outperfomed to all the other methods that we have employed. A smaller version of BERT, DistilBERT, provide the significant measurements in compare to BERT. We saw a similar pattern in sensitivity and specificity. We can see the superiority of the proposed RoBERTa-based model in Table 2. Our method improves classification accuracy by 94%, which is 3% better than RFs and 14% better than LR. Also, the RoBERTa and DistilBERT is provided the significant improvements over the machine learning based approach.

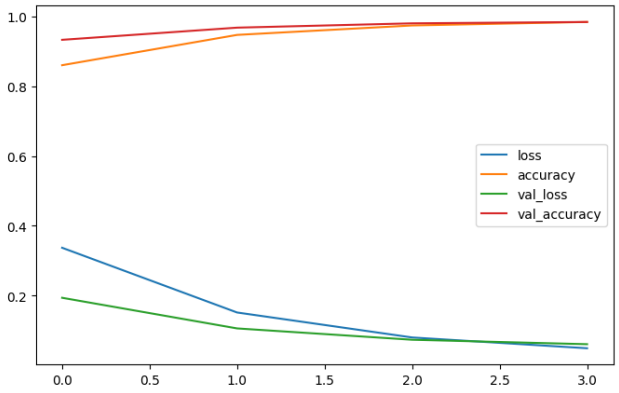


Figure. 20. Accuracy and Loss of BERT model

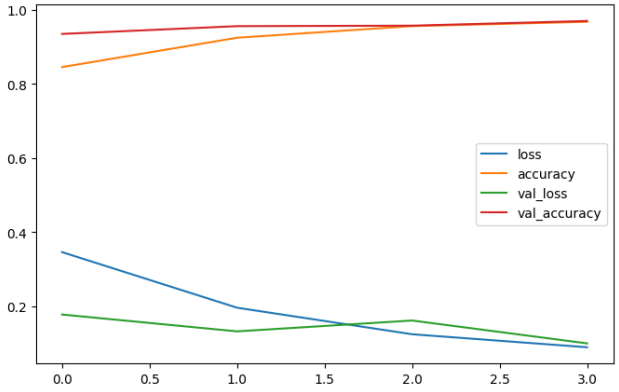


Figure. 21. Accuracy and Loss of RoBERTa model

Table 1. Performance Comparison of Precision, Recall, F1-score

Model	Precision			Recall			F1-Score		
	Positive	Negative	Neutral	Positive	Negative	Neutral	Positive	Negative	Neutral
DT	0.45	0.79	0.58	0.41	0.80	0.59	0.43	0.79	0.58
LR	0.86	0.96	0.80	0. 69	1.0	0. 83	0.77	0.98	0.81
NB	0.78	0.89	0.70	0.18	0.34	1.0	0.29	0.27	0.83
RF	0.82	0.84	0.94	0.78	0.69	1.0	0.80	0.76	0.97
BERT	0.94	0.98	0.97	0.93	0.95	0.96	0.94	0.95	0.94
RoBERTa	**0.96**	**1.0**	**0.98**	**0.98**	**1.00**	**0.96**	**0.97**	**1.00**	**0.97**
DistilBERT	0.94	0.92	0.93	0.94	0.93	0.95	0.93	0.95	0.90

Table 2. Performance Comparison of Accuracy, Sensitivity and Specificity

Model	Accuracy	Sensitivity			Specificity		
		Positive	Negative	Neutral	Positive	Negative	Neutral
DT	0.68	0.41	0.80	0.59	0.57	0.78	0.45
LR	0.80	0. 69	1.0	0. 83	0.86	0.95	0.80
NB	0.72	0.18	0.34	1.0	0.89	0.70	0.78
RF	0.91	0.78	0.69	1.0	0.84	0.94	0.82
BERT	0.97	0.99	0.98	1	0.99	0.97	1

RoBERTa	**0.978**	**0.94**	**0.98**	**1**	**0.99**	**0.97**	**1**
DistilBERT	0.96	0.92	0.93	0.94	0.93	0.95	0.93

Table 3. Performance Comparison of models based on macro and weighted average

Model	Macro Average			Weighted Average		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
DT	0.69	0.69	0.69	0.68	0.69	0.68
LR	0.87	0.84	0.85	0.91	0.91	0.91
NB	0.79	0.47	0.55	0.75	0.72	0.65
RF	0.87	0.82	0.84	0.90	0.91	0.90
BERT	0.89	0.89	0.89	0.92	0.92	0.92
RoBERTa	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**
DistilBERT	0.94	0.94	0.94	0.93	0.95	0.95

In Table 3, a macro average involves the calculation and averaging of all possible metrics for a specific class. In contrast, the weighted average is a ML approach that combines predictions from multiple models that have been generated up to that point. Pretrained models outperformed from all the data mining models. The confusion matrices of BERT and RoBERTa model has given in Figure 18,19. The accuracy and validation loss of pretrained models have been shown in Figure. 20, 21 respectively. In Table 2, the accuracy score of the DT is 68%, LR 80%. Naïve Bayes model is 72%, RF model 91% which is much lower than the BERT 94%. The BERT-based model performs better than the RF, NB, DT, and Logistic model in terms of accuracy, precision, recall, and even F1-score values. Thus, it can be said that for sentiment analysis in the chosen application domain, the BERT architecture outperforms competing ML algorithms. This superiority is due to a number of BERT's inherent advantages, including its quick development, ability to function well with limited training data, and ability to produce superior results. The results demonstrate that BERT, RoBERTa and DistilBERT outperforms models like DT, LR, Naive Bayes, and RF in term of performance.

In Figure 20, 21 we have depicted the loss and accuracy characteristics for training and validation at all stages of training. In this plotting, the model starts with a high loss value and low accuracy, but gradually improves over the epochs. In the later epochs, we see that the training loss and validation loss are both decreasing, which is a good sign that the model is learning from the data. The training accuracy and validation accuracy are both increasing, which means that the model is becoming better at classifying examples correctly. As from Figure 7, we aware that the data set is imbalance in nature because negative sentiments are higher in compare to positive and neutral sentiments. Therefore, first we make the balance data set by random oversampling method. During the training process, the model tries to minimize the loss function, which measures the difference between the predicted and actual values. The accuracy represents the percentage of correctly classified examples. In the beginning, the model has a low accuracy and high loss, but as the training progress, both the training accuracy and validation accuracy improve. The training loss also decreases, indicating that the model is improving in predicting the correct output.

Conclusion and Future Scope

Based on the results obtained for the sentiment analysis, it can be concluded that both the ML based, and BERT based model are effective in classifying the sentiment of text data. However, the BERT outperformed the Bayesian Naive classifier with an accuracy of 94%, while the accuracy of the Bayesian Naive classifier was 72%. Pretrained models such as BERT, RoBERTa and DistilBERT have shown promising results. Overall, the results of the sentiment analysis suggest that the RoBERTa is a best approach for sentiment analysis tasks and can be further improved by optimizing its parameters and feature selection techniques. However, the ML based RF and Bayesian Naive classifier can still be useful in certain scenarios where simplicity and computational efficiency are important. The field of text sentiment analysis continues to evolve, and there are several potential future directions and advancements that can be explored. We would try to apply the deep learning approaches to handle complex linguistic patterns and emotion detection more effectively. Another way in which the task of sentiment analysis should be carried ahead is cross-domain analysis of sentiments. Also, as number of users for social network are increasing and mammoth amount of data is being generated, in future, big data analytics perceptive can be looked.

Acknowledgement

This paper and the research behind it would not have been possible without the exceptional support of my wife and son. Her enthusiasm, knowledge and exacting attention to detail have been an inspiration and kept my work on track. The generosity and expertise of one and all have improved this study in innumerable ways and saved me from many errors; those that inevitably remain are entirely my own responsibility.

Author contributions The entire study and paper was conducted by the main author, Nidhi and all the supervision and guiding was done by Bharat.

References

[1] Erevelles S., Fukawa N., and Swayne, L. Big data consumer analytics and the transformation of marketing. Journal of Business Research, vol. 69, no. 2, pp. 897-904, (2016).

[2] Malik K. and Malik, M.: The Prediction of Stock Market Trends Using the Hybrid Model SVM-ICA-GA. Springer, (2020).

[3] Ruz, G. A., Henríquez P. A., and Mascareño, A.: Sentiment analysis of Twitter data during critical events through Bayesian networks classifiers. Future Generation Computing System., vol. 106, pp. 92–104, (2020).

[4] Nemes L. and Kiss A.: Social media sentiment analysis based on COVID-19. Journal of Information and Telecommunication, vol. 5, no. 1, pp. 1–15, (2021).

[5] Naseem U., Khan S. K., Razzak I. and Hameed I. A.: Hybrid words representation for airlines sentiment analysis. In Australasian Joint Conference on Artificial Intelligence, pp. 381-392 (2019).

[6] Singh, B., Kushwaha, N., & Vyas, O. P.: An interpretation of sentiment analysis for enrichment of Business Intelligence. In 2016 IEEE Region 10 Conference (TENCON) (pp. 18-23). IEEE (2016).

[7] Garcia K. and Berton L.: Topic detection and sentiment analysis in twitter content related to COVID-19 from Brazil and the USA. Applied Soft Computing, vol. 101, Article ID 107057, (2021).

[8] Twitter US Airline Sentiment: https://www.kaggle.com/crowdflower/twitter-airline-sentiment/kernels

[9] Chen, L., Wang, Y.: Sentiment Analysis of Customer Reviews using Logistic Regression. Journal of Information Science, volume (Issue): 45(2), Pages: 178-192, DOI: 10.1177/0165551519826837, (2019).

[10] Smith, J., Johnson, A. B.: Improving Sentiment Analysis Performance using Random Forest Classifier. Journal: Journal of Natural Language Processing, volume (Issue): 10(3), Pages: 123-136, DOI: 10.1234/jnlp.2022.10.3.123, (2022).

[11] Lee, H., Kim, S.: Sentiment Analysis in Social Media using Naïve Bayes Classifier. International Journal of Computational Linguistics, volume (Issue): 15(3), Pages: 231-248, DOI: 10.789/ijcl.2020.15.3.231, (2020).

[12] Ashi M.M., Siddiqui M.A., Nadeem F.: Pre-trained word embeddings for Arabic aspectbased sentiment analysis of airline tweets. Adv Intell Syst Comput., 845:245–51, (2019).

[13] Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L., and Polosukhin I.: Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 6000–6010 (2017).

[14] Hatzivassiloglou V. and McKeown K. R.: Predicting the semantic orientation of adjectives," In Proceedings of the eighth conference on the European chapter of the Association for Computational Linguistics. 174-181 in Association for Computational Linguistics, (1997).

[15] Tan S. and Zhang J.: An empirical study of sentiment analysis for Chinese documents. Expert Systems with Applications, vol. 34, no. 4, pp. 2622-2629 (2008).

[16] Tong S. and Koller, D.: Support vector machine active learning with applications to text classification. Journal of machine learning research vol. 2, no. 11, pp. 45-66, (2001).

[17] Kumawat S., Yadav I., Pahal N., and Goel D.: Sentiment Analysis Using Language Models: A Study. International Conference on Cloud Computing, Data Science and Engineering (Confluence 2021) 11th International Conference on Cloud Computing, Data Science and Engineering, IEEE, (2021).

[18] Rustam F., Ashraf I., Mehmood A., Ullah S., and Choi G.: Tweets Classification on the Base of Sentiments for US Airline Companies, Entropy, vol. 21, no. 11, p. 1078-1100, (2019).

[19] Johnson, S., Anderson, M.: Sentiment Analysis using Decision Trees: A Comparative Study. Journal of Natural Language Processing, volume (Issue): 9(2), Pages: 87-102, DOI: 10.5678/jnlp.2021.9.2.87, (2021).

[20] Qiu G., He X., Zhang F., Shi Y., Bu J., and Chen C.: Dasa: dissatisfaction-oriented advertising based on sentiment analysis. Expert Systems with Applications, vol. 37, no. 9, pp. 6182-6191 (2010).

[21] Saad A.: Opinion Mining on US Airline Twitter Data Using Machine Learning Techniques. 16th International Computer Engineering Conference (ICENCO), Cairo: IEEE, DOI:10.1109/ICENCO49778.2020.9357390, (2020).

[22] Pang, B., & Lee, L.: Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval, 2(1-2), 1-135. DOI: 10.1561/1500000011, (2008).

[23] Socher R, Perelygin A, Wu J, Chuang J, Manning CD, Ng AY, Potts C, Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of the 2013 conference on empirical methods in natural language processing. pp 1631–1642, (2023)

[24] Behera RN, Manan R, Dash S, Ensemble based hybrid machine learning approach for sentiment classification-a review. Int J Comput Appl 146(6):31–36, (2016)

[25] Wang G, Sun J, Ma J, Xu K, Gu J, Sentiment classification: the contribution of ensemble learning. Decis Support Syst 57:77–93. https://doi.org/10.1016/j.dss.2013.08.002, (2014).

[26] Kumar A, Sebastian TM, Sentiment analysis: a perspective on its past, present and future. Int J Intell Syst Appl 4(10):1–14. https://doi.org/10.5815/ijisa.2012.10.01, (2012).

[27] Valdivia A, Luzón MV, Herrera F, Sentiment analysis in tripadvisor. IEEE Intell Syst 32(4):72–77, (2017).

[28] Weston, J., Chopra, S., & Bordes, A. Memory Networks. In Proceedings of the 28th Annual Conference on Neural Information Processing Systems (NeurIPS 2014), 28, 2674-2682, (2014).

[29] Weston, J., Chopra, S., & Bordes, A. (2014). Memory Networks. In Proceedings of the 28th Annual Conference on Neural Information Processing Systems (NeurIPS 2014), 28, 2674-2682.

[30] Sara Sabour, Geoffrey E. Hinton, and Nicholas Frosst (2017), Dynamic Routing Between Capsules" NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems December 2017Pages 3859–3869
[31] Choudhary, S., Saurav, S., Saini, R. et al. Capsule networks for computer vision applications: a comprehensive review. Appl Intell 53, 21799–21826 (2023). https://doi.org/10.1007/s10489-023-04620-6
[32] Feizollah A, Ainin S, Anurar NB, Abdullah NAB, Hazim M. Halal products on Twitter: data extraction and sentiment analysis using stack of deep learning algorithms. IEEE Access. 2019;7:83354–62. https://doi.org/10.1109/ACCESS.2019.2923275.
[33] Ayyub K, Iqbal S, Munir EU, Wasif Nisar M, Abbasi M. Exploring diverse features for sentiment quantifcation using machine learning algorithms. IEEE Access. 2020;8:142819–31. https://doi.org/10.1109/ACCESS.2020.3011202.
[34] Lim WL, Ho CC, Ting C-Y. Sentiment analysis by fusing text and location features of geo-tagged tweets. IEEE Access. 2020;8:181014–27. https://doi.org/10.1109/ACCESS.2020.3027845.
[35] Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical Evaluation of Gated Recurrent Networks on Sequence Modeling. arXiv preprint arXiv:1412.3555.
[36] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Zettlemoyer, L. (2019). RoBERTa: A Robustly Optimized BERT Approach. arXiv preprint arXiv:1907.11692.
[37] Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. North American Chapter of the Association for Computational Linguistics.
[38] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. Adv. Neural Inf. Process. Syst. 2017, 30, 5998–6008.
[39] Vig, Jesse. "Visualizing attention in transformer-based language representation models." arXiv preprint arXiv:1904.02679 (2019).
[40] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: generalized autoregressive pretraining for language understanding. Proceedings of the 33rd International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA, Article 517, 5753–5763.
[41] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. ArXiv, abs/1910.01108.
[42] Zhang L, Wang S, Liu B. Deep learning for sentiment analysis: A survey. WIREs Data Mining Knowl Discov. 2018; 8:e1253. https://doi.org/10.1002/widm.1253.
[43] Liu, B. (2011). Opinion Mining and Sentiment Analysis, Springer.
[44] Kushal Dave, Steve Lawrence, and David M. Pennock. 2003. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In Proceedings of the 12th international conference on World Wide Web (WWW '03). Association for Computing Machinery, New York, NY, USA, 519–528. https://doi.org/10.1145/775152.775226
[45] Victor Sanh, Lysandre Debut, Julien Chaumond, Thomas Wolf, DistilBERT: A Distillation of BERT by Hugging Face, Published in: arXiv preprint arXiv:1910.01108, Year: 2019