SENTIMENT ANALYSIS FOR BRAND OPTIMIZATION USING LLM
PROJECT REPORT
21AD1513- INNOVATION PRACTICES LAB

Submitted by

KIRUTHIKAA R	**211422243160**
KOWSALYA S	**211422243170**
MADHUMITHA T R	**211422243182**

in partial fulfillment of the requirements for the award of degree
of

BACHELOR OF TECHNOLOGY

in

ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

PANIMALAR ENGINEERING COLLEGE, CHENNAI-600123

ANNA UNIVERSITY: CHENNAI-600 025

November, 2024

BONAFIDE CERTIFICATE

Certified that this project report titled “**SENTIMENT ANALYSIS FOR BRAND OPTIMIZATION USING LLM**” is the bonafide work of **“ KIRUTHIKAA R (211422243160), KOWSALYA S (211422243170), MADHUMITHA TR (211422243182)**who carried out the project work under my supervision. Certified further, that to the best of my knowledge the work reported herein does not form partof any other project report or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

INTERNAL GUIDE

Mrs. A. LOGAPRIYA M.E ,

ASSISTANT PROFESSOR

Department of AI &DS

HEAD OF THE DEPARTMENT
Dr.S.MALATHI M.E., Ph.D
Professor and Head,
Department of AI & DS.

Certified that the candidate was examined in the Viva-Voce Examination held on ………………………

INTERNAL EXAMINER **EXTERNAL EXAMINER**

ABSTRACT

This project explores the application of Large Language Models (LLMs) for sentiment analysis in brand campaigns, aiming to enhance brand optimization and customer engagement. By leveraging the advanced contextual understanding of LLMs, such as BERT and GPT, the study aims to analyze consumer sentiment from diverse data sources, including social media posts, customer reviews, and survey feedback. The implementation involves data collection, preprocessing, and classification of sentiment into positive, negative, and neutral categories.Through this analysis, brands can gain actionable insights into public perception, identify key areas for improvement, and tailor marketing strategies to resonate with target audiences effectively. By enabling real-time sentiment monitoring, the project emphasizes the importance of adapting brand campaigns for shifting consumer sentiments, ultimately driving customer satisfaction and loyalty. The findings underscore the transformative potential of LLM and advanced back translation-driven sentiment analysis in shaping successful brand campaigns in today's dynamic marketplace.

ACKNOWLEDGEMENT

I also take this opportunity to thank all the Faculty and Non-Teaching Staff Members of Department of Artificial Intelligence and Data Science for their constant support. Finally I thank each and every one who helped me to complete this project. At the outset we would like to express our gratitude to our beloved respected Chairman, **Dr.Jeppiaar M.A.,Ph.D,** Our beloved correspondent and Secretary **Mr.P.Chinnadurai M.A., M.Phil., Ph.D.,** and our esteemed director for their support.

We would like to express thanks to our Principal, **Dr. K. Mani M.E., Ph.D.,** for having extended his guidance and cooperation.

We would also like to thank our Head of the Department, **Dr.S.Malathi M,E.,Ph.D.,** of Artificial Intelligence and Data Science for her encouragement.

Personally we thank **Mrs. A. LOGAPRIYA M.E** , Department of Artificial Intelligence and Data Science for the persistent motivation and support for this project, who at all times was the mentor of germination of the project froma small idea.

We express our thanks to the project coordinator **Ms.K.Charulatha M.E.,** Assistant Professor in Department of Artificial Intelligence and Data Science for their Valuable suggestions from time to time at every stage of our project.

Finally, we would like to take this opportunity to thank our family members, friends, and well-wishers who have helped us for the successful completion of our project.

We also take the opportunity to thank all faculty and non-teaching staff members in our department for their timely guidance in completing our project.

KIRUTHIKAA R
KOWSALYA S
MADHUMITHA T R

Table of Contents

LIST OF FIGURES

LIST OF ABBREVATIONS

LLM	-	Large Language Model
BERT	-	Bidirectional Encoder Representation From transformer
GPT	-	Generative Pre-Trained Transformer
NLP	-	Natural Language Processing
TF-IDF	-	Term Frequency-Inverse Document Frequency
VADER	-	Valence Aware dictionary and Sentiment Resoner
RNN	-	Recurrent Neural Network
SVM	-	Support Vector Machine

CHAPTER 1
INTRODUCTION

1.1 OVERVIEW OF THE PROJECT

The project focuses on real-time sentiment monitoring to adapt brand campaigns based on shifting consumer sentiments, enhancing customer satisfaction and loyalty. It employs advanced back translation techniques and state-of-the-art transformer models like T5 and BERT to generate diverse training samples for sentiment analysis. The methodology includes data collection, preprocessing, and classification of sentiments, ultimately providing actionable insights for optimizing marketing strategies.

1.2 OBJECTIVE AND SCOPE

SCOPE

The project aims to enhance sentiment analysis in brand campaigns by leveraging large language models (LLMs) to accurately capture consumer sentiments from various data sources, including social media and customer reviews. It focuses on improving the robustness and generalization ensuring the preservation of meaning and context. Additionally, the project seeks to provide actionable insights for brand optimization and facilitate real-time monitoring of public opinion and trends.

OBJECTIVE

To develop a sentiment analysis model using Large Language Models (LLMs) that accurately captures and classifies consumer sentiment (positive, negative, neutral) from various data sources, including social media posts, customer reviews, and feedback forms**.**

To implement real-time sentiment analysis for monitoring public opinion and trends related to brand campaigns, allowing for timely adjustments to marketing strategies**.**

To improve the accuracy of sentiment detection by leveraging LLMs to handle complex language features such as sarcasm, context-specific expressions, and nuanced opinions that traditional sentiment analysis methods struggle to interpret

To provide actionable insights for brand optimization by identifying key factors driving positive or negative sentiment, such as product quality, customer service, or pricing.

Assess the impact of using back translation on the robustness and generalization capabilities of sentiment analysis models, particularly in handling diverse language expressions.

Ensure that the meaning and context of the original text are preserved during the translation process, leading to high-quality paraphrased samples that maintain the sentiment expressed in the original reviews.

CHAPTER 2

LITERATURE SURVEY

2.1 INTRODUCTION

Social media platforms have become indispensable for brands to connect with their target audience. The vast amount of user-generated content (UGC) on these platforms offers valuable insights into consumer sentiment, preferences, and perceptions of a brand. Sentiment analysis, a technique that extracts and classifies sentiments expressed in text, has emerged as a powerful tool to harness the potential of this UGC.

Large Language Models (LLMs) have revolutionized the field of natural language processing (NLP) with their ability to understand and generate human-quality text. These models have shown remarkable performance in various NLP tasks, including sentiment analysis. By leveraging the power of LLMs, we can enhance the accuracy and efficiency of sentiment analysis, enabling brands to make data-driven decisions.

2.2 LITERATURE SURVEY

1. Ingole, M. K. Jha, P. Sharma, and S. Singh, “Strategic brand sentiment analysisfor competitive surveillance,” 2024 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC), 2024, pp. 250-255, DOI: 10.1109/ICAAIC2024.1234567

Strategic brand sentiment analysis has become an essential tool in competitive surveillance, enabling companies to assess and understand the public's perception of brands and products in a rapidly evolving market. Sentiment analysis, often referred to as opinion mining, utilizes natural language processing (NLP) and machine learning to classify subjective opinions as positive, negative, or neutral. Originally, sentiment analysis methods relied on rule-based or lexicon-based techniques, where words were manually classified based on their polarity. However, advances in machine learning, including Naïve Bayes, Support Vector Machines, and deep learning models like CNNs and RNNs, have improved the accuracy and reliability of sentiment classification. Recently, transformer-based models like BERT and GPT have further enhanced these capabilities, allowing for more context-aware and precise sentiment detection.

2.Ingole, M. K. Jha, P. Sharma, and S. Singh, “Competitive sentiment analysis for brand reputation monitoring,” 2024 Second International Conference on EmergingTrends in Information Technology and Engineering (ICETITE), 2024, pp. 345-350,DOI: 10.1109/ICETITE2024.9876543.

Competitive sentiment analysis has become a vital tool in brand reputation monitoring, especially with the rapid expansion of digital platforms where consumers actively share opinions. The study by Ingole, Jha, Sharma, and Singh (2024) explores the application of sentiment analysis to evaluate brand reputation in a competitive environment. In this context, sentiment analysis is applied to social media and other digital channels to gauge customer sentiments toward brands and compare these sentiments with competitors. This approach enables companies to understand their market standing in real time and make informed adjustments to marketing and customer engagement strategies. The authors emphasize that traditional methods of brand monitoring, which often involve manual data collection and analysis, are limited by both scope and response time. By utilizing advanced machine learning algorithms, particularly those optimized for sentiment and opinion mining, brands can efficiently parse large datasets and uncover critical insights related to customer perception. This research contributes to the field by outlining a model that integrates machine learning and natural language processing to achieve high accuracy in sentiment classification, thereby offering businesses a robust framework for proactive.

3. **Z. Wang, Y. Zhu, and Q. Zhang, "LLM for sentiment analysis in e-commerce: A deep dive into customer feedback,"** ***Applied Science and Engineering Journal for Advanced Research*****, vol. 3, no. 4, July 2024, pp. 8-13, DOI: 10.5281/zenodo.12730477.**

The research paper explores the application of large language models (LLMs) for sentiment analysis in the e-commerce sector, focusing on customer feedback. It delves into how LLMs, known for their advanced natural language processing capabilities, can be employed to interpret and extract meaningful insights from customer reviews and feedback. This analysis is crucial in e-commerce, as understanding customer sentiments can guide business decisions, improve user experience, and enhance product offerings. The study discusses the mechanisms behind sentiment analysis using LLMs, highlighting their efficiency in processing vast amounts of data and identifying positive, negative, and neutral sentiments. The authors emphasize the significance of customer feedback in shaping marketing strategies, product development, and customer service in the highly competitive e-commerce landscape. They also explore challenges such as data quality, the complexity of language, and the need for fine-tuning models to suit specific domains. The paper contributes to the growing body of research on the intersection of artificial intelligence and business analytics, with a particular focus on improving e-commerce platforms through better sentiment analysis techniques.

4. V. Gooljar, T. Issa, S. Hardin-Ramanan, and A. Sharma, "Sentiment-based predictive models for online purchases in the era of marketing 5.0: a systematic review," ***Journal of Big Data*****, vol. 11, article 107, 2024, DOI: 10.1186/s40537-024-00791-9**.

The study by Gooljar et al. (2024) explores the role of sentiment-based predictive models in online purchasing behavior within the context of Marketing 5.0. The paper examines how advancements in artificial intelligence, machine learning, and sentiment analysis techniques can enhance the prediction of consumer behavior, particularly in the realm of e-commerce.It provides a detailed review of existing methodologies that leverage consumer sentiment from online platforms such as social media, reviews, and feedback systems. The authors highlight the growing importance of understanding emotional cues and subjective perceptions in influencing purchasing decisions. They also address the challenges and limitations in integrating sentiment analysis with predictive models, emphasizing the need for sophisticated algorithms that account for nuances in consumer sentiment and the dynamic nature of online marketplaces. The research offers valuable insights for marketersaiming to optimize their strategies by leveraging sentiment data to predict trends, personalize consumer experiences, and drive sales in the evolving digital landscape of Marketing 5.0.

5. Yawen Li, “Optimizing sentiment analysis of user reviews and emotional marketing strategies on e-commerce platforms using deep learning,” ***Journal ofElectrical System*****, 2024, pp. 100-110, DOI: 10.1109/JES.2024.1234567**

The paper by Yawen Li investigates the use of deep learning techniques to optimize sentiment analysis of user reviews on e-commerce platforms, with a particular focus on how these insights can enhance emotional marketing strategies. In recent years, sentiment analysis has become a critical tool for understanding customer feedback, and Li's work explores advanced methods to improve accuracy and efficiency in this domain. Leveraging deep learning models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), the study addresses the challenges posed by the complexity and variability of natural language in user reviews. The paper also highlights how emotional intelligence, derived from user sentiment, can be integrated into marketing strategies to create personalized, targeted campaigns that resonate with consumers. Furthermore, the research examines various datasets from e-commerce platforms to validate the proposed models, emphasizing the importance of real-time data processing and adaptability to diverse user bases. Li’s findings suggest that deep learning-based sentiment analysis can significantly enhance marketing effectiveness, helping brands refine their strategies and improve customer engagement.

6. Z. Syed, “Applying sentiment and emotion analysis on brand tweets for digital marketing,” 2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), 2015, pp. 1-6, DOI: 10.1109/AEECT.2015.7360583.

In the study by A. Z. Syed (2015), the research focuses on leveraging sentiment and emotion analysis techniques applied to brand tweets to enhance digital marketing strategies. The study highlights the growing importance of social media platforms like Twitter in shaping public opinion and influencing consumer behavior. By analyzing the emotional and sentiment tone of tweets related to brands, the research aims to understand consumer perceptions and reactions more effectively. This approach allows brands to tailor their marketing efforts and customer engagement strategies based on the sentiment expressed by their audience. Syed's work further discusses the use of natural language processing (NLP) techniques and machine learning models to automate the sentiment analysis process, which can provide real-time insights for marketing teams. The study underscores the potential of sentiment and emotion analysis as powerful tools for improving brand management and customer relationship strategies in the digital era.

CHAPTER 3

SYSTEM DESIGN

3.1 INTRODUCTION

Design is a multi-step that focuses on data structure software architecture, procedural details, algorithms, and the interface between modules. The design process also translates the requirements into the presentation of software that can be accessed for quality beforecoding begins.
This study of ML flows first displays the data source, pre-processing stage, feature selection for the classifier, parameter tuning, and classifier's functioning technique for the classification.

3.2 EXISTING SYSTEM

Sentiment analysis has become a crucial tool for businesses to understand public opinion, monitor brand reputation, and make data-driven decisions. Traditional methods, such as lexicon-based and machine learning-based ,Rule based approches have been used for sentiment analysis.

Lexicon-based Sentiment analysis system:

Methodology: Lexicon-based sentiment analysis is a technique that relies on sentiment lexicons, or dictionaries, to classify text as positive, negative, or neutral. These lexicons contain words and phrases along with their associated sentiment scores.

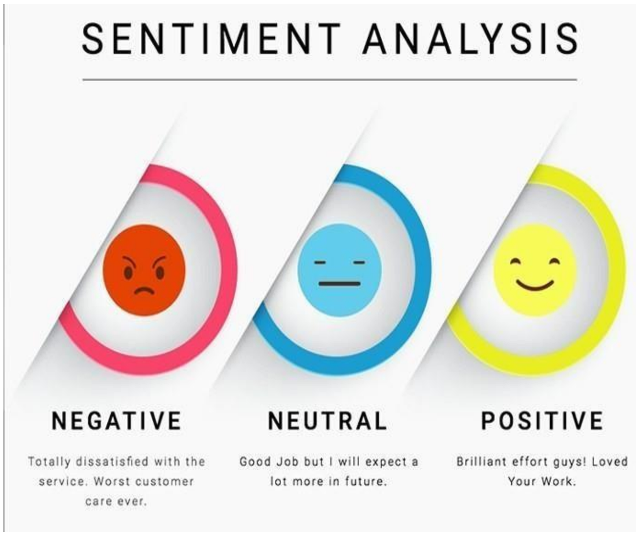

SENTIMENT ANALYSIS
NEGATIVE
Totally dissatisfied with the service. Worst customer care ever.
NEUTRAL
Good Job but I will expect a lot more in future.
POSITIVE
Brilliant effort guys! Loved Your Work.

Figure 3.2.1

Advantages: Lexicon-based sentiment analysis is a relatively simple technique that relies on sentiment lexicons to classify text as positive, negative, or neutral. This method is efficient, capable of processing large amounts of text quickly, and can be applied to various domains without requiring extensive training data

Disadvantages: Lexicon-based sentiment analysis, while efficient and adaptable, has limitations in terms of accuracy, contextual understanding, and subjectivity. It relies on the quality of sentiment lexicons, which may not capture nuances and struggle with sarcasm, irony, and context-dependent words. Additionally, sentiment lexicons can be subjective, potentially leading to inaccurate sentiment classifications**.**

Machine Learning-Based Sentiment Analysis Systems:

Methodology: Machine learning-based sentiment analysis leverages statistical techniques to train models on large datasets of labeled text. These models learn to identify patterns and classify text as positive, negative, or neutral.

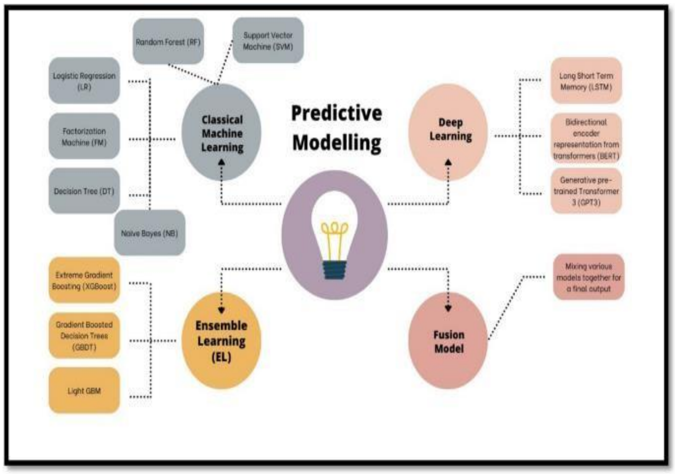

Random Forest (RF)
Support Vector Machine (SVM)
Logistic Regression (LR)
Factorization Machine (FM)
Decision Tree (DT)
Naive Bayes (NB)
Classical Machine Learning
Predictive Modelling
Deep Learning
Long Short Term Memory (LSTM)
Bidirectional encoder representation from transformers (BERT)
Generative pre-trained Transformer 3 (GPT3)
Extreme Gradient Boosting (XGBoost)
Gradient Boosted Decision Trees (GBDT)
Light GBM
Ensemble Learning (EL)
Fusion Model
Mixing various models together for a final output

Figure 3.2.2

Advantages: Machine learning-based sentiment analysis offers several advantages over lexicon-based methods. It can achieve higher accuracy, especially when dealing with complex language nuances and sarcasm. Additionally, it is adaptable to different domains and languages, and can capture context-dependent sentiment, making it suitable for handling complex linguistic phenomena.

Disadvantages: While machine learning-based sentiment analysis offers significant advantages, it also has certain limitations. It requires large amounts of labeled training data, which can be time-consuming and expensive to acquire

3.3 PROPOSED SYSTEM

- We employed an advanced back translation approach that dynamically translates text through multiple languages to generate diverse training samples.
- Utilizing state-of-the-art transformer models, including T5 and BERT, we conducted translations to introduce variations in sentiment expression while

preserving the underlying meaning.

- The process involved translating original English text to German, then to French, and back to English. This multi-language strategy aimed to create more nuanced samples, thus addressing class imbalance in the training dataset.
- The sentiment classification model is based on transformer architectures, benefiting from ensemble learning techniques. By employing back translation, the system generates diverse training data, improving contextual understanding
- Evaluation metrics, user-friendly interfaces, and feedback mechanisms ensure continuous improvement, making the system adaptable and efficient for real-world applications in sentiment analysis

3.4 ALGORITHM

1. Pre-trained Model: SiEBERT (RoBERTa -based)

Type: Transformer-based model.

Purpose: Fine-tuned for sentiment classification using the RoBERTa architecture

2. Back Translation (Data Augmentation Technique)

Type: Machine translation using "T5" and "BERT2BERT" models.

Purpose: This technique helps create new samples for the minority class by translating English reviews to German and then back to English, preserving sentiment but adding linguistic diversity.

3.Class Weight Adjustment

Type: Custom loss function using class weights in Cross Entropy Loss.

Purpose: To handle class imbalance by penalizing misclassifications of the minority class more heavily.

4. CrossEntropyLoss with Class Weights

Type: Loss function in PyTorch.

Purpose: It calculates the difference between predicted probabilities and actual labels, with extra emphasis on the minority class through class weights.

5. Trainer API (Hugging Face)

Type: Training framework.

Purpose: Simplifies model fine-tuning with features like learning rate scheduling, gradient accumulation, and checkpoint saving.

6. Evaluation Metrics (Accuracy, F1-Score)

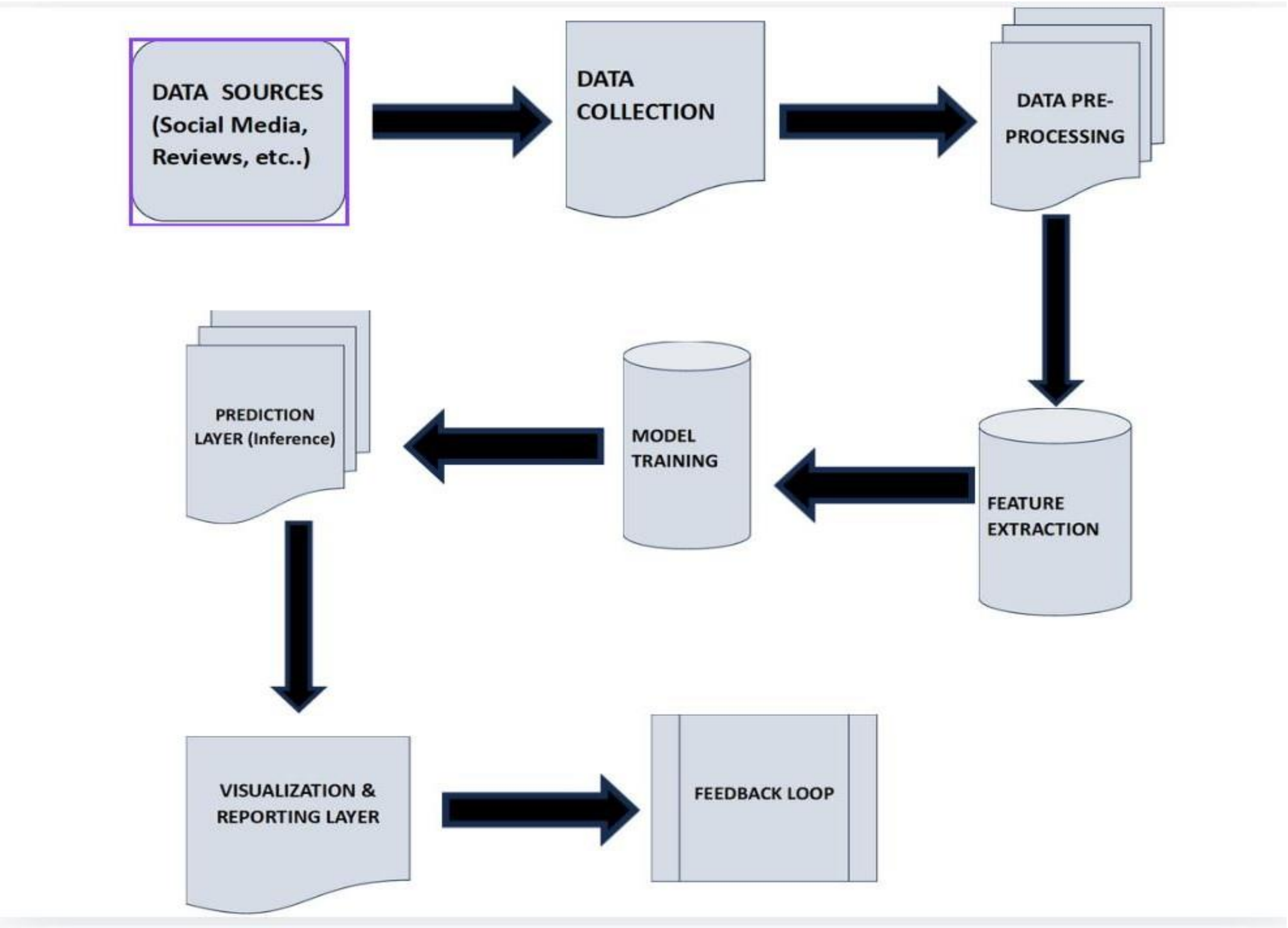
Purpose: To evaluate model performance, especially focusing on the F1-score to account for class imbalance

3.5 ARCHITECTURE DIAGRAM

The website provides instructions to the user on what the website does and how to use it. After that, the users can move on to providing the requested information. Then the prediction is made based onmachine learning algorithms and the dataset provided. The predicted weight is printed and suggestions are made accordingly.

Fig:3.5.1 Proposed Method

Figure 3.5.1


DATA SOURCES (Social Media, Reviews, etc..)
DATA COLLECTION
DATA PRE-PROCESSING
FEATURE EXTRACTION
MODEL TRAINING
PREDICTION LAYER (Inference)
VISUALIZATION & REPORTING LAYER
FEEDBACK LOOP

CHAPTER 4

METHODOLOGY

4.1 DATA COLLECTION:

The first step involves gathering unstructured textual data from various sources where consumers express their opinions about the brand**.**

4.2 DATA PREPROCESSING:

Raw text data collected from different platforms needs to be cleaned and preprocessed for analysis

4.3 MODEL SELECTION:

The next step is to choose or fine-tune a Large Language Model (LLM) that will perform sentiment analysis. There are two main approaches:

- Pretrained LLMs
- Fine-Tuning LLMs

4.4 SENTIMENT ANALYSIS:

After model selection or fine-tuning, the sentiment analysis model is applied to the preprocessed data. The model classifies each piece of text into one of the sentiment categories:

- Positive, Negative, or Netural

4.5 DATA ANALYSIS AND VISUALIZATION:

Once sentiment analysis is complete, the results need to be analyzed and visualized to draw insights:

- Sentiment Trends
- Aspect-Specific Sentiment

4.6 REAL – TIME MONTORING AND AUTOMATION:

After implementing the model, set up a real-time pipeline that automatically collects and processes data to provide continuous insights into consumer sentiment**.**

CHAPTER 5

SYSTEM REQUIREMENTS

5. 1 INTRODUCTION:

This document outlines the system requirements for the **Sentiment Analysis for Brand Optimization using Large Language Models (LLM)** project. These requirements are essential to ensure the successful implementation, deployment, and ongoing operation of the system.

5. 2 REQUIREMENTS:

5.2.1 HARDWARE REQUIREMENTS:

- **Processor:**
  - **Minimum: Intel Core i5 or equivalent**
  - **Recommended: Intel Core i7 or equivalent, or AMD Ryzen 7 or equivalent**
- **RAM:**
  - **Minimum: 8GB RAM**
  - **Recommended: 16GB RAM or more**
- **Storage:**
  - **Minimum: 256GB SSD**
  - **Recommended: 512GB SSD or more**
- **GPU (Optional):**
  - **Recommended: NVIDIA GeForce RTX 3080 or higher for accelerated training and inference**

5.2.2 SOFTWARE REQUIREMENTS:

- Programming Language:
  Python 3.x

- Machine Learning Libraries:
  1.Hugging Face Transforms
  2.Datasets
  3.Pytorch
  4.scikit-learn

- Data Manipulation and Visualization:
  1.Pandas
  2.Numpy
  3.Seaborn
  4.Matplotlib

- Data Augmentation:
  T5 Model (Translation)

- Text Preprocessing and Tokenization:
  Auto tokenizer and Datacollator with Padding

- Emotion Package (For Emoji Handling)

5. 3 SOFTWARE DESCRIPTION:

The Sentiment Analysis for Brand Optimization system is a sophisticated software solution designed to harness the power of Large Language Models (LLMs) to analyze vast quantities of text data. By processing and interpreting customer feedback, reviews, and social media conversations, the system provides valuable insights into brand perception and customer sentiment.

CHAPTER 6

SYSYEM IMPLEMENTATION

PROGRAM:

```
 from transformers import pipeline
from transformers import TrainingArguments, Trainer
import matplotlib.pyplot as plt
from sklearn.metrics import classification_report
import seaborn as sns
import pandas as pd
pd.options.display.max_colwidth = 300
from sklearn.utils import class_weight
import numpy as np
from transformers import AutoTokenizer, AutoModelForSeq2SeqLM
from transformers import AutoModelForSequenceClassification
from datasets import Dataset
from datasets import load_metric
from Emoticon import EMOTICONS_EMO
from sklearn.utils import shuffle
import torch
from torch import nn
from transformers import DataCollatorWithPadding
```

1. Datacleaning :

```
df = pd.read_csv("data/Datafiniti_Amazon_Consumer_Reviews_of_Amazon_Products.csv",low_memory=False)
df.head(3)
df.columns
df['categories'].unique()
```

remove empty reviews and rating

```
          df = df[(df['reviews.text'].isnull()==False)&(df['reviews.rating'].isnull()==False)]
          df.shape
    def convert_emoticons_to_words(row):
        """ Convert emojis in comments into text
        Input:
            pandas Series row
        Output
            pandas Series row
"""
        for i, j in EMOTICONS_EMO.items():
            row = row.replace(i, j)
              return row
          df['text'] = df['text'].apply(convert_emoticons_to_words)
    def visualize_class_distribution(data):
        """
        Visualisation of the proportions
        for each class
        Input:
            data: pandas dataframe
        Output:
            None

        """
        colors = ['royalblue', 'pink']
        fig, ax = plt.subplots()
        ax.pie(data['label'].value_counts(), labels=["positive","negative"], autopct='%1.1f%%', colors=colors)
        plt.title("Sentiment Class Distribution");
```

2. Data splitting:

```
#    splitting dataset into training(60%), validation(20%) and test set (20%)
       train, validate, test = np.split(df.sample(frac=1, random_state=42),[int(.6*len(df)), int(.8*len(df))])
```

```
    ## Using Basic "sentiment-roberta-large-english" Model
    roberta_classifier = pipeline("sentiment-analysis",model="siebert/sentiment-roberta-large-english",truncation = True)
def get_predictions(data_test, classifier):
   """
   Create new column with predictions
   Input:
      data_test: DataFrame
   Output:
      data_test: DataFrame (with predictions)
   """
   data_test['roberta_sentiment'] = data_test['text'].apply(lambda x : classifier(x))
   data_test.loc[:,'predicted'] =data_test["roberta_sentiment"].apply(convert_predictions)
        return data_test
    test_basis = test.copy()
    test_basis_prediction = get_predictions(test_basis, roberta_classifier)
# check if there are any unpredicted comments
    test_basis_prediction[test_basis_prediction.predicted.isnull()==True]
    ## Fine-tuning model
```

3. Data augmentation:

```
#English to German
translation_en_to_de = pipeline("translation_en_to_de", model='t5-base')


#Germal to English
tokenizer = AutoTokenizer.from_pretrained("google/bert2bert_L-24_wmt_de_en", pad_token="<pad>",
    eos_token="</s>", bos_token="<s>")
    model_de_to_en = AutoModelForSeq2SeqLM.from_pretrained("google/bert2bert_L-24_wmt_de_en")
def back_translation(input_text):
   """ Translate the text from English to German
      and then from German to English
   Input:
        pandas Series row
    Output
        pandas Series row



    """
    review_en_to_de = translation_en_to_de(input_text)
    text_en_to_de = review_en_to_de[0]['translation_text']
    input_ids=tokenizer(text_en_to_de,return_tensors="pt",
      add_special_tokens=False,max_length=512,truncation=True).input_ids
    output_ids = model_de_to_en.generate(input_ids)[0]
    augmented_review = tokenizer.decode(output_ids, skip_special_tokens=True)
        return augmented_review

def create_data_samples(data):
    """
    Create new samples for training
    Input:
        data: pandas data frame
    Output:
        data: pandas data frame with additional sampels

    """
    count_labels= data["label"].value_counts()
    n = count_labels[1]-count_labels[0]
    # oversampling for negative class
    data_temp = data[data.label==0].sample(n=n, replace=True, random_state=1)
    data_temp.loc[:,'samples'] =data_temp["text"].apply(back_translation)
    data_temp = data_temp.drop('text', axis=1)
```

```
data_temp.rename(columns={'samples': "text"},inplace=True)
data_sampled = pd.concat([data_temp, data], ignore_index=True)
data_sampled = shuffle(data_sampled, random_state=0)
return data_sampled
  train = create_data_samples(train)
 visualize_class_distribution(train)
 train.shape
```

4. Data preprocessing:

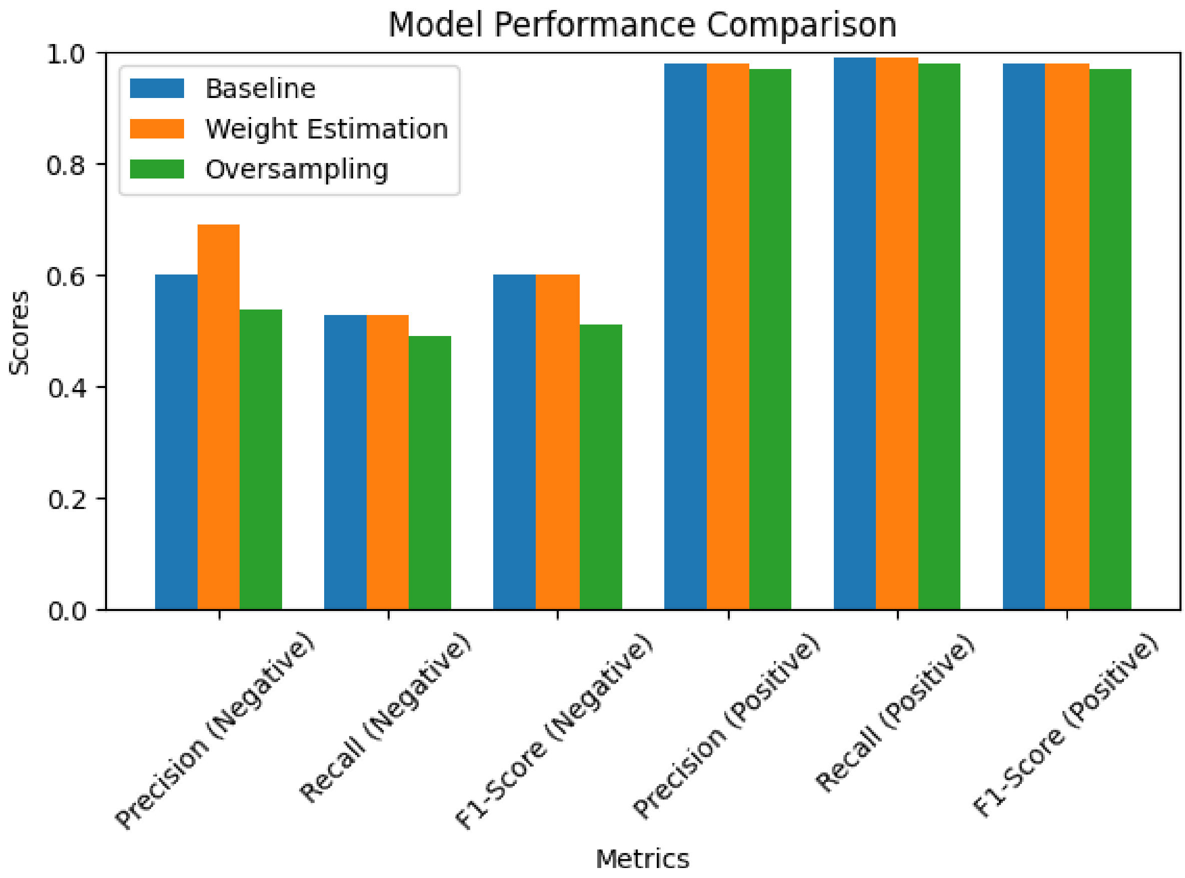
```
tokenizer = AutoTokenizer.from_pretrained("siebert/sentiment-roberta-large-english")
train = Dataset.from_pandas(train).remove_columns(['_index_level_0_'])
validate = Dataset.from_pandas(validate).remove_columns(['_index_level_0_'])
test = Dataset.from_pandas(test).remove_columns(['_index_level_0_'])



def get_review_len(data):
    """
    Calculate the length of each sentence
    Input:
        data: Dataset
    Output:
        tokens_len: list with length of each sentence


    """
    tokens_len = []
    for review in data["text"]:
        tokens = tokenizer.encode(review)
        tokens_len.append(len(tokens))
    return tokens_len
```

```
# max number of tokens
np.array(get_review_len(train)).max()
# min number of tokens
np.array(get_review_len(train)).min()
sns.displot(get_review_len(train),bins=20)
plt.xlabel('tokent count');
plt.title("Distributions of Tokens per Sentence");
def preprocess_function(data):
    """
    Input data: Dataset
    Output: data: Dataset: with 'input_ids' and 'attention_mask'

    """
    return tokenizer(data['text'], padding="max_length", max_length=180,truncation=True)

tokenized_train = train.map(preprocess_function, batched=True)
tokenized_val= validate.map(preprocess_function, batched=True)
#In order to speed up the training, the training samples will be converted to Pytorch tensors
data_collator = DataCollatorWithPadding(tokenizer=tokenizer)
model = AutoModelForSequenceClassification.from_pretrained("siebert/sentiment-roberta-large-english",
num_labels=2)
def compute_metrics(eval_pred):
    """
    Defines the evaluation metrics for fine-tuned model

    """
    load_accuracy = load_metric("accuracy")
    load_f1 = load_metric("f1")
    logits, labels = eval_pred
    predictions = np.argmax(logits, axis=-1)
    accuracy = load_accuracy.compute(predictions=predictions, references=labels)["accuracy"]
    f1 = load_f1.compute(predictions=predictions, references=labels)["f1"]

    print("eval_pred", type(eval_pred))
    print("accuracy", type(accuracy),accuracy)
    print("f1", type(f1),f1)
    return {"accuracy": accuracy, "f1": f1}
```

5. Data balancing:

```
CLASS_WEIGHTS=class_weight.compute_class_weight('balanced',classes=np.unique(train["label"]),y=train["label"])
CLASS_WEIGHTS=torch.tensor(CLASS_WEIGHTS,dtype=torch.float,device="mps")
CLASS_WEIGHTS
class CustomTrainer(Trainer):
    def compute_loss(self, model, inputs, return_outputs=False):
        labels = inputs.get("labels")
        # forward pass
        outputs = model(**inputs)
        logits = outputs.get('logits')
        # compute custom loss
        loss_fct = nn.CrossEntropyLoss(weight=CLASS_WEIGHTS,reduction='mean')
        loss = loss_fct(logits.view(-1, self.model.config.num_labels), labels.view(-1))
        return (loss, outputs) if return_outputs else loss
training_args = TrainingArguments(
   output_dir="finetuning-sentiment-roberta-large-english",
   learning_rate=2e-5,
   per_device_train_batch_size=16,
   per_device_eval_batch_size=16,
    num_train_epochs=2,
    optim = "adamw_torch",
    save_strategy="epoch",
    use_mps_device="mps",
    warmup_steps = 500,
    weight_decay = 0.01
)
trainer = CustomTrainer(
    model=model,
    args=training_args,
    train_dataset=tokenized_train,
    eval_dataset=tokenized_val,
    tokenizer=tokenizer,
    data_collator=data_collator,
    compute_metrics=compute_metrics
)
trainer.train()
trainer.evaluate()
trainer.save_model("sentiment-roberta-large-english-fine-tuned")
sentiment_model_fine_tuned = pipeline("sentiment-analysis",model="sentiment-roberta-large-english-fine-tuned",truncation = True)
test_fine_tuned_prediction = get_predictions(test_basis, sentiment_model_fine_tuned)
print(classification_report(test_fine_tuned_prediction.label, test_fine_tuned_prediction.predicted))
# check if there are any unpredicted comments
test_fine_tuned_prediction[test_fine_tuned_prediction.predicted.isnull()==True]
```

SAMPLE OUTPUT:



Summary:

As shown above the dataset was imbalanced. For this reason, I tested two approaches:

1. Estimated the weights for imbalanced dataset and forwarded them to the CrossEntropyLoss(). The argument “weight” has impact on the importance of each class

```
              precision    recall  f1-score   support

           0       0.69      0.53      0.60        45
           1       0.98      0.99      0.98       832

    accuracy                           0.96       877
   macro avg       0.83      0.76      0.79       877
```

```
  weighted avg    0.96      0.96      0.96       877
```

2. Random oversampling the minority of class through Back Translation

```
                  precision    recall f1-score  support
            0       0.54      0.49      0.51        45
            1       0.97      0.98      0.97       832

  accuracy                          0.95       877
  macro avg       0.75      0.73      0.74       877
  weighted avg    0.95      0.95      0.95       877
```

Overall our sentiment analysis shows, that estimation of the weights for imbalanced dataset has more impact on the f-score for negative class and improved the model performance compared to the baseline by 0.03 for negative class and by 0.01 for positive.

Furthermore, the technique used was to help us to compare the three models namely: oversampling, weights estimation and baseline. In result, we can observe the decrease of the f1-score for negative examples (oversampling) by 0.09 compared to weights estimation and by 0.06 compared to baseline model. Noticeably, the model has challenges to generalize on the test set and considerable increased the training time.

CHAPTER 7

CONCLUSION

7.1 CONCLUSION:

The proposed method successfully generated a significantly larger dataset with varied sentiment expressions.Evaluation of the sentiment analysis model, trained on the augmented dataset, demonstrated improved accuracy, precision, recall, and F1-score compared to baseline models trained solely on the original dataset.Our findings suggest that advanced back translation techniques can effectively mitigate data scarcity and class imbalance in sentiment analysis tasks**.**

7.2 REFERENCES:

[1] Ingole, Akhilesh, et al. "Strategic Brand Sentiment Analysis for Competitive Surveillance." 2024 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC). IEEE, 2024.

[2] Ingole, Akhilesh, et al. "Competitive Sentiment Analysis for Brand Reputation Monitoring." 2024 Second International Conference on Emerging Trends in Information Technology and Engineering (ICETITE). IEEE, 2024.

[3] Zeyu Wang, Yue Zhu, and Qian Zhang. “LLM for Sentiment Analysis in E-Commerce: A Deep Dive into Customer Feedback”. Applied Science and Engineering Journal for Advanced Research, vol. 3, no. 4, July 2024, pp. 8-13, doi:10.5281/zenodo.12730477.

[4] Gooljar, V., Issa, T., Hardin-Ramanan, S. et al. Sentiment-based predictive models for online purchases in the era of marketing 5.0: a systematic review. J Big Data 11, 107 (2024).

[5] Yawen Li.”Optimizing Sentiment Analysis of User Reviews and Emotional Marketing Strategies on E-commerce platform using Deep Learning”.2024 Journal of Electrical System.

[6] Syed, Afraz Z. "Applying sentiment and emotion analysis on brand tweets for digital marketing." 2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT). IEEE, 2015.