MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE

Deemed to be University

(Declared under Distinct Category by Ministry of Education, Govt. of India)

NAAC Accredited with A++ Grade

Project Report

On

Development of Sentimental analysis Chatbot

Submitted By:

Akansha Singh Rajawat

(0901CA221005)

Industry Mentor:

Mr. Tushar, Project coordinator (Technook Edutech)

Faculty Mentor:

Dr. R.S. Jadon , Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE

Gwalior – 474005(MP) estd.1957

Jan – june 2024

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE

Deemed to be University

(Declared under Distinct Category by Ministry of Education, Govt. of India)

NAAC Accredited with A++ Grade

Project Report

On

Development of Sentimental analysis Chatbot

A project report submitted in partial fulfillment of the requirement for the degree of

MASTER IN COMPUTER APPLICATION

In

COMPUTER SCIENCE AND ENGINEERING

Submitted By:

Akansha Singh Rajawat

(0901CA221005)

Industry Mentor:

Mr. Tushar, Project coordinator (Technook Edutech)

Faculty Mentor:

Dr. R.S. Jadon , Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE

Gwalior – 474005(MP) estd.1957

Jan – June 2024

teachnook

Certification of internship completion.

Dear,

AKANSHA SINGH RAJAWAT

We are heartful to announce the enrollment for an internship. And we are grateful to announce that he has gone through an internship in the domain of Artificial Intelligence in the months of Jan.
We are in collaboration with **Cognizance 24 IIT ROORKEE**. while working students have gathered commendable soft and hard skills.

Enrollment month:- Jan 15th to March 25th

Yours Faithfully,

SAUMYA TIWARI
5167, 9th Main Rd,
Sector 6, HSR Layout,
Bengaluru, Karnataka 560102

LinkedIn TOP 2023 STARTUPS

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE

Deemed to be University

(Declared under Distinct Category by Ministry of Education, Govt. of India)

NAAC Accredited with A++ Grade

CERTIFICATE

This is to certified that Ms. **Akansha Singh Rajawat (0901CA221005)** has submitted the project report titled **Development of Sentimental analysis Chatbot** under the mentorship of **Mr. Tushar (**Project Coordinator Technook Edutech **),** in partial fulfilment of the requirement for the award of degree of **master in computer Application**, submitted in Department of Computer Science and Engineering, **Madhav Institute of Technology and Science, Gwalior**.

26/4/24

Dr. R.S Jadon

(professor and Project Coordinator)

Computer Science and Engineering

26/4/24

Dr. Manish Dixit

(Professor and head)

Computer Science and Engineering

Dr. Manish Dixit
Professor & HOD
Department of CSE
M.I.T.S. Gwalior

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE

Deemed to be University

(Declared under Distinct Category by Ministry of Education, Govt. of India)

NAAC Accredited with A++ Grade

DECLARATION

I hereby declare that the work being presented in this project report, for the partial fulfillment of requirement for the award of the degree of Master in Computer Application in Computer Science and Engineering at **Madhav Institute of Technology & Science, Gwalior** is an authenticated and original record of my work under the mentorship of **Mr. Tushar, Project Coordinator, Technook Edutech.**

Akansha Singh Rajawat

0901CA221005

2022-2024

Master in Computer Application

Computer Science and Engineering

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE

Deemed to be University

(Declared under Distinct Category by Ministry of Education, Govt. of India)

NAAC Accredited with A++ Grade

DECLARATION

I hereby declare that the work being presented in this project report, for the partial fulfillment of requirement for the award of the degree of Master in Computer Application in Computer Science and Engineering at **Madhav Institute of Technology & Science, Gwalior** is an authenticated and original record of my work under the mentorship of **Mr. Tushar, Project Coordinator, Technook Edutech.**

Akansha Singh Rajawat

0901CA221005

2022-2024

Master in Computer Application

Computer Science and Engineering

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE

Deemed to be University

(Declared under Distinct Category by Ministry of Education, Govt. of India)

NAAC Accredited with A++ Grade

ACKNOWLEDGEMENT

The full semester project has proved to be pivotal to my career. I am thankful to my institute, **Madhav Institute of Technology and Science** to allow me to continue my disciplinary project. I extend my gratitude to the Director of the institute, **Dr. R. K. Pandit** and Dean Academics, **Dr. Manjaree Pandit** for this.

I would sincerely like to thank my department, **Department of Computer Science and Engineering**, for allowing me to explore this project. I humbly thank **Dr. Manish Dixit**, Professor and Head, Department of Computer Science and Engineering, for his continued support during the course of this engagement, which eased the process and formalities involved.

I would like to extend my heartfelt appreciation to Mr. Tushar, Project Coordinator, Technook Edutech for their exceptional mentorship, guidance, and assistance throughout the project. Their valuable inputs and feedback have helped me enhance my knowledge and skills. Their constant encouragement and support have been instrumental in the successful completion of this project.

I am sincerely thankful to my faculty coordinator. I am grateful to the guidance of **Dr. R.S. Jadon** Professor, Computer Science and Engineering, for his continued support and guidance throughout the project. I am also very thankful to the faculty and staff of the department.

Akansha Singh Rajawat

0901CA221005

2022-2024

Master in Computer Application

Computer Science and Engineering

ABSTRACT

Sentiment Analysis also known as Opinion Mining refers to the use of natural language processing, text analysis to systematically identify, extract, quantify, and study affective states and subjective information. Sentiment analysis is widely applied to reviews and survey responses, online and social media, and healthcare materials for applications that range from marketing to customer service to clinical medicine.

In recent years, sentiment analysis has emerged as a powerful tool for understanding and analyzing human emotions expressed in textual data. Leveraging this technology, we propose the development of a sentiment analysis chatbot aimed at enhancing user experience in various domains. This chatbot utilizes state-of-the-art natural language processing techniques to analyze user input and accurately determine the sentiment conveyed.

The primary objective of the sentiment analysis chatbot is to provide users with personalized and empathetic responses based on their expressed sentiments. By understanding the emotions behind user messages, the chatbot can tailor its interactions to be more relevant, supportive, and engaging. Additionally, the chatbot can identify and address negative sentiments in real-time, allowing for timely intervention and resolution of user concerns.

Key features of the proposed sentiment analysis chatbot include sentiment classification, emotion detection, and sentiment-aware responses. Through sentiment classification, the chatbot categorizes user messages into positive, negative, or neutral sentiments, enabling it to adapt its responses accordingly. Emotion detection capabilities enable the chatbot to recognize nuanced emotions such as joy, sadness, anger, and fear, further enhancing its ability to empathize with users.

Furthermore, the sentiment analysis chatbot integrates sentiment-aware responses that are tailored to the user's emotional state. Whether expressing happiness, frustration, or sadness, the chatbot responds with empathy and understanding, fostering a more positive user experience. Additionally, the chatbot can escalate issues flagged as negative sentiments to human operators for further assistance, ensuring comprehensive support for users.

The proposed sentiment analysis chatbot has wide-ranging applications across various industries, including customer service, healthcare, education, and marketing. By harnessing the power of sentiment analysis, organizations can improve customer satisfaction, identify areas for improvement, and build stronger relationships with their audience.

In conclusion, the development of a sentiment analysis chatbot represents a significant advancement in natural language processing technology, offering enhanced user experiences and improved communication in diverse domains. Through its ability to understand and respond to user sentiments, the chatbot holds promise for revolutionizing human-computer interactions and fostering more empathetic and supportive digital environments.

सार

सेंटीमेंट एनालिसिस को ओपिनियन माइनिंग के रूप में भी जाना जाता है, जो प्रभावशाली स्थितियों और व्यक्तिपरक जानकारी को व्यवस्थित रूप से पहचानने, निकालने, मात्रा निर्धारित करने और अध्ययन करने के लिए प्राकृतिक भाषा प्रसंस्करण, पाठ विश्लेषण के उपयोग को संदर्भित करता है। भावना विश्लेषण व्यापक रूप से समीक्षाओं और सर्वेक्षण प्रतिक्रियाओं, ऑनलाइन और सोशल मीडिया, और विपणन से लेकर ग्राहक सेवा से लेकर नैदानिक चिकित्सा तक के अनुप्रयोगों के लिए स्वास्थ्य देखभाल सामग्री पर लागू होता है।

हाल के वर्षों में, पाठ्य डेटा में व्यक्त मानवीय भावनाओं को समझने और उनका विश्लेषण करने के लिए भावना विश्लेषण एक शक्तिशाली उपकरण के रूप में उभरा है। इस तकनीक का लाभ उठाते हुए, हम विभिन्न डोमेन में उपयोगकर्ता अनुभव को बढ़ाने के उद्देश्य से एक भावना विश्लेषण चैटबॉट के विकास का प्रस्ताव करते हैं। यह चैटबॉट उपयोगकर्ता इनपुट का विश्लेषण करने और व्यक्त की गई भावना को सटीक रूप से निर्धारित करने के लिए अत्याधुनिक प्राकृतिक भाषा प्रसंस्करण तकनीकों का उपयोग करता है।

भावना विश्लेषण चैटबॉट का प्राथमिक उद्देश्य उपयोगकर्ताओं को उनकी व्यक्त भावनाओं के आधार पर व्यक्तिगत और सहानुभूतिपूर्ण प्रतिक्रियाएँ प्रदान करना है। उपयोगकर्ता संदेशों के पीछे की भावनाओं को समझकर, चैटबॉट अपनी बातचीत को अधिक प्रासंगिक, सहायक और आकर्षक बना सकता है। इसके अतिरिक्त, चैटबॉट वास्तविक समय में नकारात्मक भावनाओं को पहचान और संबोधित कर सकता है, जिससे समय पर हस्तक्षेप और उपयोगकर्ता की चिंताओं का समाधान हो सकता है।

प्रस्तावित भावना विश्लेषण चैटबॉट की मुख्य विशेषताओं में भावना वर्गीकरण, भावना का पता लगाना और भावना-जागरूक प्रतिक्रियाएँ शामिल हैं। भावना वर्गीकरण के माध्यम से, चैटबॉट उपयोगकर्ता संदेशों को सकारात्मक, नकारात्मक या तटस्थ भावनाओं में वर्गीकृत करता है, जिससे वह अपनी प्रतिक्रियाओं को तदनुसार अनुकूलित करने में सक्षम हो जाता है। भावनाओं का पता लगाने की क्षमताएं चैटबॉट को खुशी, उदासी, क्रोध और भय जैसी सूक्ष्म भावनाओं को पहचानने में सक्षम बनाती हैं, जिससे उपयोगकर्ताओं के साथ सहानुभूति रखने की उसकी क्षमता और बढ़ जाती है।

इसके अलावा, भावना विश्लेषण चैटबॉट भावना-जागरूक प्रतिक्रियाओं को एकीकृत करता है जो उपयोगकर्ता की भावनात्मक स्थिति के अनुरूप होते हैं। चाहे खुशी, निराशा या दुख व्यक्त करना हो, चैटबॉट सहानुभूति और समझ के साथ प्रतिक्रिया करता है, जिससे अधिक सकारात्मक उपयोगकर्ता अनुभव को बढ़ावा मिलता है। इसके अतिरिक्त, चैटबॉट उपयोगकर्ताओं के लिए व्यापक समर्थन सुनिश्चित करते हुए, आगे की सहायता के लिए मानव ऑपरेटरों को नकारात्मक भावनाओं के रूप में चिह्नित मुद्दों को बढ़ा सकता है।

प्रस्तावित भावना विश्लेषण चैटबॉट में ग्राहक सेवा, स्वास्थ्य देखभाल, शिक्षा और विपणन सहित विभिन्न उद्योगों में व्यापक अनुप्रयोग हैं। भावना विश्लेषण की शक्ति का उपयोग करके, संगठन ग्राहकों की संतुष्टि में सुधार कर सकते हैं, सुधार के क्षेत्रों की पहचान कर सकते हैं और अपने दर्शकों के साथ मजबूत संबंध बना सकते हैं।

निष्कर्ष में, एक भावना विश्लेषण चैटबॉट का विकास प्राकृतिक भाषा प्रसंस्करण प्रौद्योगिकी में एक महत्वपूर्ण प्रगति का प्रतिनिधित्व करता है, जो विभिन्न डोमेन में बेहतर उपयोगकर्ता अनुभव और बेहतर संचार प्रदान करता है। उपयोगकर्ता की भावनाओं को समझने और उन पर प्रतिक्रिया देने की अपनी क्षमता के माध्यम से, चैटबॉट मानव-कंप्यूटर इंटरैक्शन में क्रांति लाने और अधिक सहानुभूतिपूर्ण और सहायक डिजिटल वातावरण को बढ़ावा देने का वादा करता है।

LIST OF CONTENTS

CHAPTER 1: INTRODUCTION

Sentiment is an attitude, thought, or judgment prompted by feeling. Sentiment analysis, which is also known as opinion mining, studies people's sentiments towards certain entities. From a user's perspective, people are able to post their own content through various social media, such as forums, micro-blogs, or online social networking sites. From a researcher's perspective, many social media sites release their application programming interfaces (APIs), prompting data collection and analysis by researchers and developers. However, those types of online data have several flaws that potentially hinder the process of sentiment analysis. The first flaw is that since people can freely post their own content, the quality of their opinions cannot be guaranteed. he second flaw is that ground truth of such online data is not always available. A ground truth is more like a tag of a certain opinion, indicating whether the opinion is positive, negative, or neutral.

In the ever-evolving landscape of artificial intelligence and natural language processing, the integration of sentiment analysis has opened up new avenues for enhancing user interactions and experiences. Sentiment analysis, a subfield of natural language processing, focuses on discerning the emotions, opinions, and attitudes expressed within textual data. Leveraging this technology, the development of sentiment analysis chatbots represents a significant advancement in the realm of conversational agents.

Chatbots have become ubiquitous across various industries, serving as virtual assistants, customer support agents, and companions in digital spaces. However, traditional chatbots often lack the ability to comprehend and respond appropriately to the emotional nuances conveyed by users. This limitation can result in impersonal interactions and missed opportunities to provide empathetic support.

Recognizing the importance of understanding user sentiments, the integration of sentiment analysis into chatbot systems offers a transformative solution. By analyzing the sentiment embedded within user messages, sentiment analysis chatbots can tailor their responses to align with the emotional context of the conversation. This capability enables chatbots to provide more personalized, empathetic, and effective interactions with users, ultimately enhancing user satisfaction and engagement.

The objective of this paper is to explore the potential of sentiment analysis chatbots in revolutionizing human-computer interactions across various domains. Through a comprehensive examination of sentiment analysis techniques, chatbot architectures, and real-world applications, we aim to highlight the benefits and challenges associated with integrating sentiment analysis into chatbot systems.

This paper is structured as follows: first, we provide an overview of sentiment analysis and its role in natural language processing. Next, we delve into the architecture and design considerations for sentiment analysis chatbots, including sentiment classification algorithms and emotion detection techniques. We then discuss the applications of sentiment analysis chatbots across different industries, ranging from customer service to mental health support.

In conclusion, we underscore the significance of sentiment analysis chatbots in fostering more empathetic and responsive human-computer interactions. By understanding and responding to user sentiments, these chatbots have the potential to revolutionize the way we interact with technology, leading to more meaningful and satisfying user experiences.

"It is a quite boring movie…….. but the scenes were good enough. "

The given line is a movie review that states that "it" (the movie) is quite boring but the scenes were good. Understanding such sentiments require multiple tasks.

Hence, SENTIMENTAL ANALYSIS is a kind of text classification based on Sentimental Orientation (SO) of opinion they contain. Sentiment analysis of product reviews has recently become very popular in text mining and computational linguistics research.

- Firstly, evaluative terms expressing opinions must be extracted from the review.
- Secondly, the SO, or the polarity, of the opinions must be determined.
- Thirdly, the opinion strength, or the intensity, of an opinion should also be determined.
- Finally, the review is classified with respect to sentiment classes, such as Positive And Negative, based on the SO of the opinions it contains

1.1 BACKGROUND INFORMATION

Sentiment Analysis Model: This is the heart of your chatbot. You'll need a robust sentiment analysis model that can accurately classify text into positive, negative, or neutral sentiments. Deep learning models like recurrent neural networks (RNNs), convolutional neural networks (CNNs), or transformer-based models (like BERT) are commonly used for this purpose.

Training Data: Your sentiment analysis model needs to be trained on a diverse dataset that includes text samples with labeled sentiments. This dataset should cover a wide range of topics and writing styles to ensure the model generalizes well.

Natural Language Processing (NLP) Pipeline: You'll need to implement an NLP pipeline to preprocess the text data before feeding it into the sentiment analysis model. This may include tasks such as tokenization, stop word removal, stemming or lemmatization, and part-of-speech tagging.

Chatbot Interface: Design a user-friendly interface for interacting with the chatbot. This could be a web-based interface, a messaging platform integration (like Facebook Messenger or Slack), or a standalone application.

Integration with External APIs: Depending on your requirements, you may want to integrate your chatbot with external APIs to fetch or analyze data from sources like social media platforms, news websites, or customer reviews.

Deployment: Once your chatbot is developed, you'll need to deploy it to a server or cloud platform so that it can be accessed by users. This may involve containerization using Docker or deployment on platforms like AWS, Google Cloud, or Microsoft Azure.

Monitoring and Maintenance: Continuously monitor the performance of your chatbot and update it as needed. This may involve retraining the sentiment analysis model with new data, improving the NLP pipeline, or adding new features based on user feedback.

By focusing on these key components, you can create a sentimental analysis chatbot that effectively analyzes the sentiment of text input and provides appropriate responses or actions based on the detected sentiment.

1.2PROBLEM DEFINATION

In the realm of conversational AI, traditional chatbots often struggle to comprehend and appropriately respond to the emotional nuances conveyed by users. This limitation poses several challenges and impediments to providing satisfactory user experiences:

Impersonal Interactions: Conventional chatbots typically operate based on predefined rules or keyword matching techniques, leading to generic and impersonal responses. As a result, users may feel

disconnected or unsatisfied with the interaction, especially when expressing emotions or seeking empathetic support.

Misinterpretation of User Sentiments: Without the capability to analyze user sentiments, chatbots may misinterpret the underlying emotions conveyed in user messages. This can lead to inappropriate responses or misunderstandings, further exacerbating user frustration or dissatisfaction.

Ineffective Problem Resolution: In scenarios where users express negative sentiments or concerns, traditional chatbots may fail to provide adequate support or resolution. Lacking the ability to empathize or understand the emotional context, these chatbots may offer generic responses or escalate issues inefficiently, leading to unresolved user issues.

Limited Personalization: Traditional chatbots often lack the ability to personalize interactions based on user emotions or preferences. Without sentiment analysis capabilities, chatbots cannot adapt their responses to align with the user's emotional state or provide tailored support, diminishing the overall user experience.

Missed Opportunities for Engagement: Emotional engagement plays a crucial role in user interactions, influencing satisfaction and retention. Without the ability to recognize and respond to user sentiments, chatbots may miss opportunities to engage users effectively, leading to decreased user engagement and retention rates.

Addressing these challenges requires the development and integration of sentiment analysis capabilities into chatbot systems. By leveraging sentiment analysis techniques, chatbots can accurately perceive and respond to user emotions, leading to more empathetic, personalized, and effective interactions. However, implementing sentiment analysis in chatbots presents its own set of complexities and considerations, including algorithm selection, data preprocessing, and model training, which must be carefully addressed to ensure optimal performance and usability.

1.3 OBJECTIVE OF STUDY

Understanding User Sentiment: By analyzing the sentiment of user messages, chatbots can better understand the emotional state of the user. This understanding enables them to tailor responses accordingly, providing more empathetic and personalized interactions.

Improving User Experience: Sentiment analysis helps in enhancing the overall user experience by enabling chatbots to respond appropriately to positive, negative, or neutral sentiments. This ensures that users feel understood and valued, leading to increased satisfaction and engagement.

Adapting Responses: By: identifying the sentiment of user messages, chatbots can adapt their responses to align with the user's emotional state. For instance, if a user expresses frustration or dissatisfaction, the chatbot can offer support or solutions to address their concerns effectively.

Gauging Customer Satisfaction: Sentiment analysis can be used to gauge customer satisfaction levels by analyzing the sentiment of feedback or reviews received by the chatbot. This information can be valuable for businesses to identify areas for improvement and make data-driven decisions.

Monitoring Brand Perception: Chatbots equipped with sentiment analysis capabilities can monitor social media conversations and online mentions to assess the overall sentiment towards a brand or product. This enables businesses to proactively manage their online reputation and address any negative sentiments before they escalate.

Training Chatbots: Sentiment analysis can also be used to train chatbots to recognize and respond to a wide range of emotions expressed by users. By exposing chatbots to diverse datasets containing labeled sentiments, they can learn to accurately interpret and respond to different emotional cues.

Overall, the study of sentiment analysis in chatbots aims to create more emotionally intelligent and responsive conversational agents that can better serve user needs and improve overall user satisfaction.

1.4 STRUCTURAL OVERVIEW

User Interface: This is the interface through which users interact with the chatbot. It can be a web interface, a mobile app, or integration with messaging platforms like Facebook Messenger or Slack.

Natural Language Processing (NLP): NLP is the backbone of the chatbot, enabling it to understand and process natural language input from users. Within NLP, the following components are essential:

Tokenization: Breaking down user input into individual words or tokens.

Part-of-Speech Tagging: Identifying the grammatical parts of speech of each token.

Named Entity Recognition (NER): Identifying and categorizing named entities such as people, places, and organizations mentioned in the user's input.

Dependency Parsing: Analyzing the grammatical structure of sentences to understand the relationships between words.

Word Embeddings: Representing words in a high-dimensional vector space to capture semantic similarities and differences.

Sentiment Analysis Module: This module analyzes the sentiment of user input. It typically involves the following steps:

Text Preprocessing: Cleaning and preparing text data for analysis, which may include removing stop words, punctuation, and special characters, as well as stemming or lemmatization.

Feature Extraction: Extracting relevant features from the preprocessed text data, such as word frequencies or word embeddings.

Sentiment Classification: Classifying the sentiment of the text as positive, negative, or neutral. This classification can be performed using machine learning algorithms such as Support Vector Machines (SVM), Naive Bayes, or deep learning models like Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs).

Response Generation: Based on the sentiment analysis results, the chatbot generates an appropriate response. Responses can be predefined templates for common sentiments (e.g., "I'm sorry to hear that" for negative sentiment), or they can be dynamically generated based on the context and sentiment of the user's input.

Knowledge Base or Backend Integration: The chatbot may be connected to a knowledge base or backend systems to retrieve relevant information or perform tasks based on user queries or requests. This integration enables the chatbot to provide more accurate and helpful responses.

Feedback Mechanism: A feedback mechanism allows users to provide feedback on the chatbot's responses, which can be used to improve its performance over time through iterative learning and refinement.

Analytics and Monitoring: This component tracks the performance of the chatbot, including user interactions, sentiment distribution, response accuracy, and other metrics. Analytics data can be used to identify areas for improvement and optimize the chatbot's performance continuously.

By integrating these components, a sentiment analysis chatbot can effectively analyze and respond to user input, providing a more personalized and emotionally intelligent conversational experience.

1.5 SCOPE AND SIGNIFICANCE

The scope and significance of sentiment analysis are vast and impactful across various domains:

Business and Marketing: Sentiment analysis helps businesses understand customer opinions, sentiments, and preferences regarding products, services, and brands. It enables companies to gauge customer satisfaction, identify emerging trends, and make data-driven decisions to improve marketing strategies, product development, and customer experience.

Social Media Monitoring: Sentiment analysis is widely used to monitor social media platforms for public opinions, reactions, and trends. It helps organizations track brand sentiment, detect potential crises, engage with customers, and manage their online reputation effectively.

Customer Service and Support: Sentiment analysis enhances customer service by automatically categorizing and prioritizing incoming messages based on sentiment. It allows businesses to identify and address customer issues, complaints, or feedback promptly, leading to improved customer satisfaction and loyalty.

Market Research: Sentiment analysis is a valuable tool for market research, enabling researchers to analyze large volumes of textual data from surveys, reviews, forums, and social media. It provides insights into consumer behavior, preferences, and sentiment shifts, helping businesses stay competitive and adapt to changing market dynamics.

Brand Monitoring and Reputation Management: Sentiment analysis helps organizations monitor mentions of their brand, products, or services across various online platforms. It enables them to assess brand sentiment, detect sentiment trends, and take proactive measures to manage their brand reputation effectively.

Political Analysis and Public Opinion: Sentiment analysis is used in political analysis to gauge public opinion, sentiment, and reactions towards political candidates, parties, policies, and issues. It helps political organizations, policymakers, and analysts understand voter sentiment, anticipate electoral outcomes, and shape communication strategies.

Healthcare and Pharma: In healthcare, sentiment analysis is used to analyze patient feedback, reviews, and social media discussions related to healthcare providers, treatments, medications, and medical devices. It helps healthcare organizations improve patient satisfaction, identify areas for improvement, and monitor public health sentiments and concerns.

Financial Trading and Investment: Sentiment analysis is applied in financial markets to analyze news, social media discussions, and other textual data for sentiment signals that may impact stock prices, market trends, and investment decisions. It helps traders and investors make informed decisions and manage risks more effectively.

Customer Feedback Analysis: Sentiment analysis automates the analysis of customer feedback from various sources such as surveys, reviews, and feedback forms. It helps businesses extract valuable insights, identify recurring themes, and prioritize action items to address customer needs and improve overall satisfaction.

Overall, sentiment analysis plays a crucial role in extracting actionable insights from textual data, enabling organizations to make informed decisions, enhance customer experiences, manage reputational risks, and stay competitive in today's data-driven world. Its scope and significance continue to expand as more industries recognize its value in driving business outcomes and improving stakeholder engagement.

CHAPTER 2: LITERATURE REVIEW

A review of literature on sentiment analysis chatbots provides insights into the development, applications, methodologies, and challenges within this field. Here's a structured outline of what such a review might cover:

1. Introduction to Sentiment Analysis and Chatbots:

Sentiment analysis and chatbots are two fascinating fields within natural language processing (NLP) that have gained significant attention in recent years due to their practical applications in various industries. Let's break down each of them:

Sentiment Analysis:

Sentiment analysis, also known as opinion mining, is the process of determining the emotional tone behind a piece of text. It involves analyzing the text to classify it as positive, negative, or neutral based on the expressed sentiment. This analysis can be performed at different levels: document level, sentence level, or aspect level.

Applications of sentiment analysis include:

- Social media monitoring: Analyzing public opinions about products, brands, or events on platforms like Twitter, Facebook, etc.

- Customer feedback analysis: Understanding customer sentiments towards products or services through reviews, surveys, or feedback forms.

- Market research: Gauging public sentiment towards certain topics or trends to make informed business decisions.

- Brand reputation management: Monitoring online discussions to identify and address any negative sentiment towards a brand.

Sentiment analysis techniques include machine learning algorithms, lexicon-based methods, and deep learning models like recurrent neural networks (RNNs) or transformers.

2. Chatbots:

A chatbot is a computer program designed to simulate conversation with human users, typically over the internet. Chatbots can be rule-based or AI-powered. Rule-based chatbots follow predefined rules and patterns to respond to user inputs, while AI-powered chatbots leverage natural language understanding and machine learning techniques to provide more dynamic and contextually relevant responses.

Applications of chatbots include:

- Customer service: Assisting customers with inquiries, troubleshooting, or providing product information.

- Virtual assistants: Helping users perform tasks like setting reminders, scheduling appointments, or finding information.

- E-commerce: Guiding users through the shopping process, recommending products, and processing orders.

- Healthcare: Providing basic medical advice, scheduling appointments, or answering frequently asked questions.

Chatbots can be deployed on various platforms such as websites, messaging apps, and voice assistants, and they are often integrated with backend systems to access relevant data and services.

Conclusion: Sentiment analysis and chatbots are powerful tools that leverage NLP techniques to understand and interact with human language in meaningful ways. By analyzing sentiment, organizations can gain valuable insights into customer opinions and market trends, while chatbots enable seamless communication and automation of various tasks, leading to enhanced user experiences and operational efficiency.

2. Historical Overview:

The historical development of sentiment analysis and chatbots has been marked by significant milestones, technological advancements, and increasing adoption across various industries. Here's an overview:

Early Developments (1950s-1990s):

- Sentiment analysis traces its roots back to the 1950s, with initial attempts focusing on analyzing text for emotional content.

- Early chatbots like ELIZA, developed by Joseph Weizenbaum in the 1960s, employed simple pattern-matching techniques to simulate conversation.

- Sentiment analysis in this era primarily relied on handcrafted rules and linguistic patterns.

Emergence of Machine Learning (2000s-2010s):

- The 2000s witnessed the rise of machine learning techniques in sentiment analysis, with researchers exploring algorithms like Support Vector Machines (SVM) and Naive Bayes.

- Sentiment lexicons, such as SentiWordNet and LIWC, were developed to provide labeled data for training machine learning models.

- Chatbots evolved from rule-based systems to more sophisticated approaches, leveraging natural language understanding and generation techniques.

- Platforms like Pandorabots and AIML (Artificial Intelligence Markup Language) enabled developers to create more interactive chatbots.

3. Deep Learning Revolution (2010s-present):

- The advent of deep learning, particularly recurrent neural networks (RNNs) and later transformer models like BERT and GPT, revolutionized sentiment analysis and chatbots.

- Deep learning models showed superior performance in sentiment analysis tasks, surpassing traditional machine learning approaches.

- Chatbots became more contextually aware and capable of generating human-like responses, thanks to advances in natural language processing (NLP) and conversational AI.

- Companies started integrating sentiment analysis into chatbots to personalize interactions and tailor responses based on user sentiment.

4. Industry Adoption and Integration (2010s-present):

- Sentiment analysis and chatbots found widespread applications across industries, including marketing, customer service, healthcare, finance, and e-commerce.

- Social media platforms leveraged sentiment analysis to monitor user sentiment, analyze trends, and personalize content.

- Chatbots became integral to customer support operations, automating routine inquiries, and enhancing user engagement.

- Integration of sentiment analysis with chatbots enabled more empathetic and contextually relevant interactions, improving user satisfaction and brand perception.

5. Ethical and Societal Implications:

- As sentiment analysis and chatbots became more pervasive, concerns emerged regarding privacy, bias, and the ethical use of AI.

- Issues such as algorithmic bias, data privacy, and the potential for manipulation raised questions about responsible deployment and regulation.

- Efforts to address these concerns led to the development of guidelines, frameworks, and regulations aimed at promoting ethical AI practices.

In summary, the historical evolution of sentiment analysis and chatbots reflects a journey from early rule-based systems to sophisticated deep learning models, driving advancements in language understanding and human-computer interaction. These technologies continue to evolve, shaping the way we communicate, interact, and understand human behavior in the digital age..

By conducting a comprehensive review of literature, researchers can gain a deeper understanding of the current state-of-the-art, identify research gaps, and contribute to advancing knowledge in sentiment analysis chatbots.

After many researches, I found a dataset of 1578612 tweets in English coming from two sources: Kaggle and Sentiment140. It is composed of four columns that are ItemID, Sentiment, SentimentSource and Sentiment Text. We are only interested by the Sentiment column corresponding to our label class taking a binary value, 0 if the tweet is negative, 1 if the tweet is positive and the SentimentText columns containing the tweets in a raw format.

	2401	Borderlands	Positive	im getting on borderlands and i will murder you all ,
0	2401	Borderlands	Positive	I am coming to the borders and I will kill you...
1	2401	Borderlands	Positive	im getting on borderlands and i will kill you ...
2	2401	Borderlands	Positive	im coming on borderlands and i will murder you...
3	2401	Borderlands	Positive	im getting on borderlands 2 and i will murder ...
4	2401	Borderlands	Positive	im getting into borderlands and i can murder y...
...	...	...	...	...
74676	9200	Nvidia	Positive	Just realized that the Windows partition of my...
74677	9200	Nvidia	Positive	Just realized that my Mac window partition is ...
74678	9200	Nvidia	Positive	Just realized the windows partition of my Mac ...
74679	9200	Nvidia	Positive	Just realized between the windows partition of...
74680	9200	Nvidia	Positive	Just like the windows partition of my Mac is l...

74681 rows × 4 columns

Example of twitter posts annotated with their corresponding sentiment, 0 if it is negative, 1 if it is positive.

In the showing the first ten twitter posts we can already notice some particularities and difficulties that we are going to encounter during the preprocessing steps.

- The presence of acronyms"bf" or more complicated "APL".

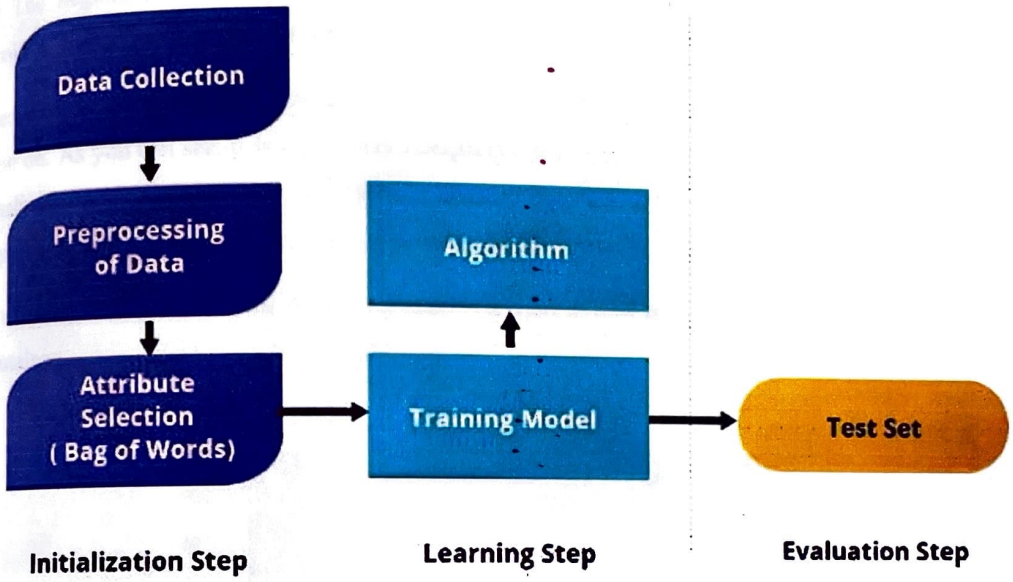
Does it means apple ? Apple (the company) ? In this context we have "friend" after so we could think that he refers to his smartphone and so Apple, but what about if the word "friend" was not here ?

- The presence of sequences of repeated characterssuch as "Juuuuuuuuuuuuuuuussssst", "hmmmm". In general when we repeat several characters in a word, it is to emphasize it, to increase its impact.

- The presence of emoticons, ":O", "T_T", ":|" and much more, give insights about user's moods.

CHAPTER 3: PROPOSED METHODOLOGY

Proposing a methodology for sentiment analysis involves outlining the steps and techniques you plan to use to analyze and classify the sentiment expressed in textual data. Here's a comprehensive methodology:



3.1. DATA COLLECTION:

To gather the data many options are possible. In some previous paper researches, they built a program to collect automatically a corpus of tweets based on two classes, "positive" and "negative", by querying Twitter with two type of emoticons:

- Happy emoticons, such as ":)", ":P", ":)" etc.
- Sad emoticons, such as ":(", ":'(", "=(".

Others make their own dataset of tweets my collecting and annotating them manually which very long and fastidious

Additionally, to find a way of getting a corpus of tweets, we need to take of having a balanced data set, meaning we should have an equal number of positive and negative tweets, but it needs also to be large enough. Indeed, more the data we have, more we can train our classifier and more the accuracy will be

• Spelling mistakesand “urban grammar” like "imgunna" or "mi".

• The presence of nounssuch as "TV", "New Moon".

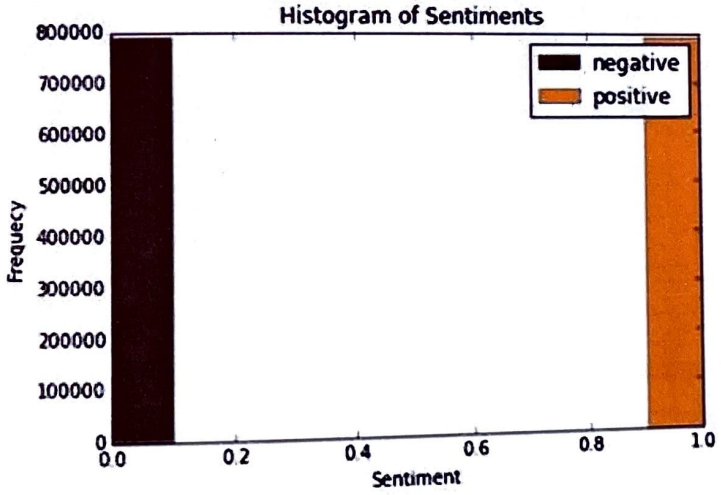
Furthermore, we can also add,

• People also indicate their moods, emotions, states, between two such as, **\cries, hummin, sigh.**

• The negation, “can't”, “cannot”, “don't”, “haven't” that we need to handle like: “I don’t like chocolate”, “like” in this case is negative.

We could also be interested by the grammar structure of the tweets, or if a tweet is subjective/objective and so on. As you can see, it is **extremely complex**to deal with languages and even more when we want to analyse text typed by users on the Internet because people 4 don’t take care of making sentences that are grammatically correct and use a ton of acronyms and words that are more or less english in our case.

We can visualize a bit more the dataset by making a chart of how many positive and negative tweets does it contains,



Histogram of the tweets according to their sentiment

We have exactly 790177 positive tweets and 788435 negative tweets which signify that the dataset is well-balanced. There are also no duplicates.

Finally, let’s recall the Twitter terminology since we are going to have to deal with in the tweets:

• Hashtag: A hashtag is any word or phrase immediately preceded by the # symbol. When you click on a hashtag, you’ll see other Tweets containing the same keyword or topic.

• @username: A username is how you're identified on Twitter, and is always preceded immediately by the @ symbol. For instance, Katy Perry is @katyperry.

• MT: Similar to RT (Retweet), an abbreviation for "Modified Tweet." Placed before the Retweeted text when users manually retweet a message with modifications, for example shortening a Tweet.

• Retweet: RT, A Tweet that you forward to your followers is known as a Retweet. Often used to pass along news or other valuable discoveries on Twitter, Retweets always retain original attribution.

• Emoticons: Composed using punctuation and letters, they are used to express emotions concisely, ";) :) ...". Now we have the corpus of tweets, we need to use other resources to make easier the preprocessing step.

3.2 FEATURE SELECTION

Feature analysis is essential for understanding the characteristics of the data and identifying the most informative aspects that contribute to sentiment classification accuracy. In the context of this sentiment analysis dataset focused on Twitter messages and entities, several features can be examined to enhance the model's performance.

1. Textual Features:

- Bag-of-Words (BoW): Constructing a BoW representation can capture the frequency of words in each message. This can be further enhanced with techniques like TF-IDF to weigh the importance of words.

- Word Embeddings: Utilizing pre-trained word embeddings like Word2Vec, GloVe, or fastText can capture semantic relationships between words and their contextual meanings.

- n-grams: Considering sequences of words rather than just individual words can capture more nuanced linguistic patterns.

2. Entity-specific Features:

- Entity Mention: Identifying whether the entity is mentioned in the message and extracting its context can provide valuable cues about sentiment.

- Entity Sentiment: Incorporating sentiment scores of the entity itself can be indicative of the sentiment expressed in the message.

- Entity Frequency: Examining how often the entity appears in the dataset and its distribution across sentiments can offer insights into its overall sentiment polarity.

3. Semantic Features:

- Sentiment Lexicons: Integrating sentiment lexicons like SentiWordNet or AFINN can provide additional context to determine the sentiment of the message.

- Semantic Role Labeling (SRL): Analyzing the roles of words in a sentence (e.g., subject, object) can help understand the sentiment expressed towards the entity.

4. Syntactic Features:

- Part-of-Speech (POS) Tagging: Analyzing the grammatical structure of sentences can provide information about sentiment-bearing words and their roles.

- Dependency Parsing: Identifying syntactic relationships between words can aid in understanding the sentiment flow within the message.

5. Contextual Features:

- Temporal Information: Considering when the message was posted can be crucial, as sentiments can vary over time.

- User Information: Incorporating user metadata such as followers count, tweet history, or verified status can add context to the sentiment expressed.

6. Additional Features:

- Emoticons and Emoji: Emoticons and emoji can convey sentiment and adding features to capture their presence and meaning can enhance classification accuracy.

- Text Length: Longer messages may contain more detailed sentiments, while shorter messages may be more concise but still carry sentiment.

By thoroughly analyzing and incorporating these features into the sentiment analysis model, it can better understand and interpret the sentiment expressed towards entities in Twitter messages, ultimately improving classification accuracy.

3.3 DATA PRE-PROCESSING

During the Preprocessing phase of our machine learning pipeline, we address the issue of missing or inconsistent items in our dataset, which is obtained from credible sources such as Kaggle and other reliable platforms. The preservation of data completeness and integrity is of utmost importance, and in order to do this, a range of strategies are utilized to eliminate null and missing values. Through a meticulous process of eliminating any discrepancies from the dataset, we establish a robust basis for subsequent analysis.

Label and other encoding for categorical columns is a fundamental technique employed in the preprocessing stage. The procedure entails the conversion of categorical variables into numerical representations, which aids in the convergence of the model and mitigates certain algorithmic biases that may emerge due to variations in feature sizes. By encoding categorical variables, we are able to enhance the accuracy of our predictive analytics by enabling the model to efficiently read and analyze these features.

An additional crucial stage in the preprocessing stage entails the scaling of features with varying ranges. The process of normalizing guarantees that all features provide an equal contribution to the computations of the model, irrespective of their initial magnitude. The process of scaling features serves to educe the influence of outliers and fluctuations in the distribution of data, hence enhancing the performance and interpretability of the model.

During the preprocessing phase, a range of functions and techniques are employed to modify and ready the data set or the purpose of training the model. Imputation and encoding functions are crucial tools in our preprocessing toolset. Imputation substitutes missing values with estimated estimates based on surrounding data, while encoding functions like one-hot encoding are used for categorical variables. Furthermore, the utilization of scaling functions like as Min-

Max scaling or Standard scaling serves to normalize the ranges of features, hence guaranteeing uniformity in the behavior of the model across diverse datasets.

In general, the aforementioned preprocessing methods play a vital role in the preliminary stages, augmenting the accuracy and reliability of our predictive analytics. Through a rigorous process of cleaning and modifying the dataset, we establish the necessary foundation for constructing a resilient classification model aimed at predicting cardiovascular illness. This endeavor serves to enhance the effectiveness of healthcare analytics and ultimately enhance patient outcomes.

The aforementioned preparation procedures in Python can be executed through the utilization of diverse libraries, including Pandas, Scikit-learn, and NumPy. The following is a concise summary of the application of these approaches, along by their respective syntax:

Certainly, let's delve into each of these data preprocessing steps:

a. Renaming column:

- Renaming columns involves giving more descriptive names to the columns in your dataset. This step is crucial for clarity and consistency, especially when working with multiple datasets or sharing your data with others. For example, if your dataset has a column named "text" that contains the Twitter messages, you might rename it to "message" for better understanding.

b. Checking for null values and inconsistent data:

- Null values are missing data points within the dataset. Checking for null values is essential to ensure data quality and avoid errors during analysis. Inconsistent data refers to values that do not adhere to the expected format or range. For instance, in a sentiment analysis dataset, a null value in the "message" column would render the data unusable for analysis. Similarly, inconsistent data like numerical values in a categorical column (e.g., sentiment labels) should be addressed.

c. Removing inconsistency:

- Once null values and inconsistent data are identified, they need to be handled appropriately. This can involve various strategies such as imputation for null values or correcting/rejecting inconsistent data points. For instance, if there are inconsistent sentiment labels (e.g., "pos" instead of "Positive"), you might standardize them to maintain consistency across the dataset.

d. Label encoding:

- Label encoding is a process of converting categorical labels into numerical representations. In sentiment analysis, the sentiment classes (e.g., Positive, Negative, Neutral) are categorical labels that need to be encoded into numerical values for model training. For example, Positive may be encoded as 1, Negative as 2, and Neutral as 3. This allows machine learning algorithms to work with categorical data more effectively.

These preprocessing steps are essential for ensuring data quality, consistency, and compatibility with machine learning algorithms. They lay the foundation for further analysis and modeling, ultimately leading to more accurate and reliable results in tasks like sentiment analysis.

has context menu

	id	country	Label	text
0	2401	Borderlands	Positive	I am coming to the borders and I will kill you...
1	2401	Borderlands	Positive	im getting on borderlands and i will kill you ...
2	2401	Borderlands	Positive	im coming on borderlands and i will murder you...
3	2401	Borderlands	Positive	im getting on borderlands 2 and i will murder ...
4	2401	Borderlands	Positive	im getting into borderlands and i can murder y...

3.4 EDA

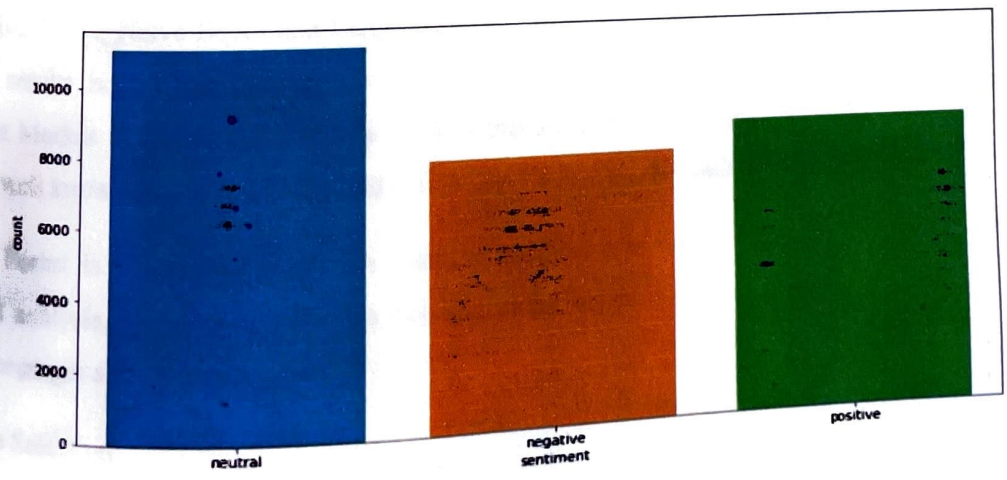
Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods.

EDA helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions.

EDA is primarily used to see what data can reveal beyond the formal modeling or hypothesis testing task and provides a provides a better understanding of data set variables and the relationships between them. It can also help determine if the statistical techniques you are considering for data analysis are appropriate. Originally developed by American mathematician John Tukey in the 1970s, EDA techniques continue to be a widely used method in the data discovery process today.

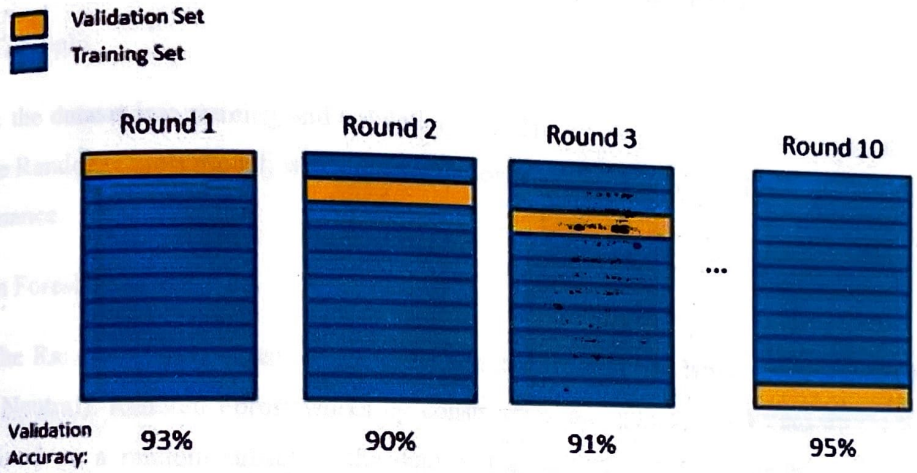
Presents the statistical description of data variable, providing summary statistics such for ach variable in the dataset

	sentiment	text
1	neutral	11117
2	positive	8582
0	negative	7781



Distribution of dataset into variables

This figure shows the graphical representation of the number of variables count in a data set



3.5 Machine Learning Modeling :

Once we have applied the different steps of the preprocessing part, we can now focus on the machine learning part. There are three major models used in sentiment analysis to classify a sentence into positive or negative: SVM, Naive Bayes and Language Models (NGram). SVM is known to be the model giving the best results but in this project, we focus only on probabilistic model that are Naive Bayes and Language Models that have been widely used in this field. Let's first introduce the Naive Bayes model which is well known for its simplicity and efficiency for text classification.

Random Forest is a powerful ensemble learning method commonly used for classification tasks like sentiment analysis. Here's how Random Forest can be applied to analyze sentiment in the context of Twitter messages and entities:

1. Feature Selection:

- Before training the Random Forest model, it's crucial to select relevant features from the dataset. These features can include textual features (like Bag-of-Words, word embeddings), entity-specific features (entity mention, sentiment, frequency), semantic features (sentiment lexicons, SRL), syntactic features (POS tagging, dependency parsing), and contextual features (temporal information, user data).

2. Data Preprocessing:

- Preprocess the data by cleaning the text, removing noise (such as URLs, special characters, and stopwords), tokenizing, and vectorizing the text using techniques like TF-IDF or word embeddings.

3. Train-Test Split:

- Divide the dataset into training and validation sets. The training set (e.g., twitter_training.csv) is used to train the Random Forest model, while the validation set (e.g., twitter_validation.csv) is used to evaluate its performance.

4. Random Forest Training:

- Train the Random Forest classifier using the selected features and labels (sentiment classes: Positive, Negative, Neutral). Random Forest works by constructing multiple decision trees during training. Each tree is trained on a random subset of the data and a random subset of features, which helps reduce overfitting and improve generalization.

5. Hyperparameter Tuning:

- Perform hyperparameter tuning to optimize the performance of the Random Forest model. Parameters like the number of trees in the forest, maximum depth of the trees, and minimum number of samples required to split a node can significantly impact the model's performance.

6. Model Evaluation:

- Evaluate the trained Random Forest model on the validation set using appropriate evaluation metrics such as accuracy, precision, recall, and F1-score. In this case, the top-1 classification accuracy is specified as the metric, which measures the proportion of correctly predicted sentiments for each message-entity pair.

7. Model Interpretation:

- Analyze the feature importances provided by the Random Forest model to understand which features contribute the most to sentiment classification. This analysis can provide insights into the key factors driving sentiment towards entities in Twitter messages.

8. Iterative Improvement:

- Iterate on the feature selection, preprocessing steps, and model training process to improve the Random Forest model's performance further. Experiment with different combinations of features and hyperparameters to find the optimal configuration.

By following these steps, the Random Forest method can effectively analyze sentiment in Twitter messages about specific entities, providing valuable insights into public opinion and sentiment trends.

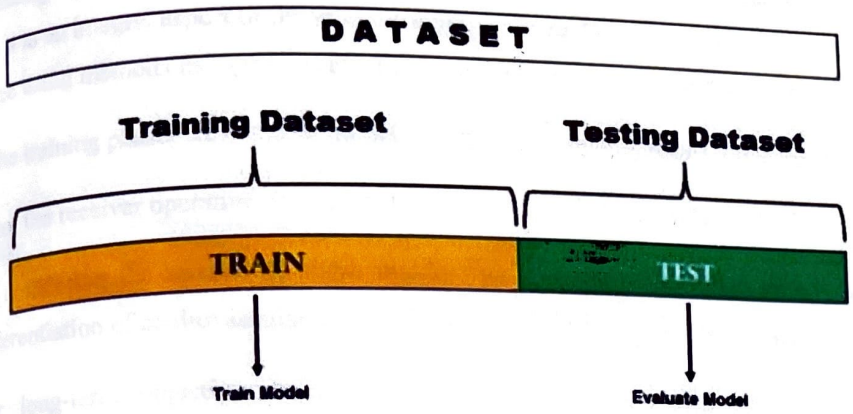
3.6 Data Splitting:

Training set and a testing set are created from the preprocessed dataset using techniques such as train-test split or cross-validation. The models are trained using the training set, and their performance is evaluated using the testing set.

This training approach applies the chosen algorithms to the training data in order to learn about patterns and correlations between characteristics and the aim variable, which in this case is the diagnosis of sentimental analysis chat bot. To avoid overfitting and maximize model performance, hyperparameter tuning is a must-have tool. Following training, models are valuated using a number of performance indicators, including F1-score, accuracy, precision, recall, and AUC-ROC. We can find out where the models are falling short in their prediction of sentimental analysis diagnoses by doing this evaluation.

We will put the most effective model into action after we determine it through the evaluations. Machine Learning's Modeling phase is crucial for developing reliable predictive models that can help detect and treat cardiovascular disease at an early stage.

Various machine learning methods used to classify sentimental analysis chatbot are briefly explained below.



Data splitting into training and test

3.7 Model Training:

In the Model Training stage of our cardiovascular disease (SENTIMENTAL ANALYSIS CHATBOT) classification challenge, we use the preprocessed dataset to train machine earning models. The primary goal of this training method is to predict the presence of SENTIMENTAL ANALYSIS CHATBOT in patients. Now is the time to optimize the settings of various classification algorithms so that they produce the best possible projected accuracy by aligning them with the training data.

Several distinct classification algorithms that perform admirably on binary classification tasks are initially chosen. The methods that fall under this category include random forests, decision trees, logistic regression, and support vector machines (SVMs). Cardiovascular disease SENTIMENTAL ANALYSIS CHATBOT research can be approached in various ways, each with its own advantages and disadvantages. Once we've decided on the algorithms, we train them using the training data subset. Models train to increase prediction accuracy by modifying internal parameters as they learn from input features and the labels that correspond to them. The goal of this iterative approach is to enhance the prediction accuracy of the models by making them better able to differentiate between individuals with and without cardiovascular disease SENTIMENTAL ANALYSIS CHATBOT.

Selecting appropriate values for the model-controlling parameters is known as hyperparameter tuning, and it is an integral aspect of the model-training process. By systematically examining the hyperparameter space using methods like grid search or randomized search, optimal parameter values can be found.

In the training phase, we evaluate model's accuracy, precision, recall, Fl-score, and area

under the receiver operating characteristic curve (AUC-ROC) in relation to the training data.

By comparing the models on these metrics, we can see how well they handle the classification and differentiation of cardiovascular disease SENTIMENTAL ANALYSIS CHATBOT) situations.

Our long-term objective is to create reliable classifiers that can accurately forecast patients' SENTIMENTAL ANALYSIS CHATBOT outcomes by training and optimizing several models with the training data. The recurrent training process establishes the framework for developing a strong prediction model that can contribute the early diagnosis and management of cardiovascular disease.

Testing and evaluating a sentiment analysis model on the entity-level dataset from Twitter involves several key steps to ensure its accuracy and effectiveness. Here's a comprehensive guide on how to perform model testing and evaluation:

1. Data Preparation: Start by loading the training dataset (`twitter_training.csv`) and the validation dataset (`twitter_validation.csv`). Preprocess the data by cleaning the text, handling null values, and encoding the labels according to the sentiment classes (Positive, Negative, Neutral). Split the data into features (messages) and labels (sentiments).

2. Feature Engineering: Extract relevant features from the text data that can help the model understand the sentiment expressed towards the entities in the messages. These features can include Bag-of-Words representations, TF-IDF vectors, word embeddings, or any other suitable text representation technique. Additionally, consider incorporating entity-specific features such as mentions, sentiment scores, and frequency.

3. Model Training: Select a suitable machine learning algorithm for sentiment analysis. Random Forest, Support Vector Machines (SVM), and Neural Networks are common choices. Train the selected model on the training dataset using the extracted features and corresponding labels. Adjust hyperparameters as needed through techniques like cross-validation to optimize performance.

4. Model Testing: After training the model, evaluate its performance on the validation dataset. Use the trained model to predict the sentiment of the messages in the validation set. Compare the predicted labels with the ground truth labels to assess the model's accuracy.

5. Evaluation Metrics: Calculate various evaluation metrics to gauge the model's performance. In this case, since the task is to predict the sentiment of Twitter messages about entities, the primary evaluation metric is top-1 classification accuracy. Additionally, consider computing precision, recall, and F1-score for each sentiment class to get a more detailed understanding of the model's performance.

6. Error Analysis: Conduct error analysis to identify patterns of misclassifications and understand where the model struggles. Look for common themes or characteristics in misclassified messages. This analysis can provide insights into areas for improvement, such as the need for better feature representation or more sophisticated modeling techniques.

7. Cross-Validation: Perform cross-validation to ensure the model's generalization ability. Split the training data into multiple folds, train the model on each fold, and evaluate its performance on the remaining folds. This technique helps to mitigate issues like overfitting and provides a more reliable estimate of the model's performance.

8. Fine-Tuning and Iteration: Based on the evaluation results and error analysis, fine-tune the model by adjusting hyper parameters, feature engineering techniques, or even trying different algorithms. Iterate this process until satisfactory performance is achieved on the validation dataset.

9. Final Evaluation and Reporting: Once the model is fine-tuned and performs well on the validation dataset, evaluate its performance on a separate test dataset, if available. Report the final evaluation metrics, including accuracy, precision, recall, and F1-score, along with any insights gained from the error analysis.

By following these steps, you can thoroughly test and evaluate the sentiment analysis model on the Twitter dataset, ensuring its reliability and effectiveness in predicting sentiment towards entities in Twitter messages.

Random Forest:

Random Forest Classifier is a popular ensemble learning algorithm used for classification tasks. It operates by constructing a multitude of decision trees during the training phase and outputs the mode of the classes (classification) or mean prediction (regression) of the individual trees. Here's a breakdown of how Random Forest Classifier works:

1. Ensemble of Decision Trees: Random Forest is composed of a collection of decision trees, where each tree is trained independently on a random subset of the training data. These decision trees collectively form the "forest."

2. Random Subset Selection: During the training phase of each decision tree, a random subset of the training data is sampled with replacement (bootstrapping). This process is known as bagging (Bootstrap Aggregating). Additionally, a random subset of features is selected at each node of the tree to consider for splitting. This ensures diversity among the individual trees and reduces overfitting.

3. Decision Tree Training: Each decision tree in the Random Forest is trained recursively using a subset of the training data. At each node of the tree, the algorithm selects the best split among the random subset of features based on a criterion such as Gini impurity or information gain. This process continues until the tree reaches a specified maximum depth or no further improvement can be made.

4. Voting for Classification: Once all decision trees are trained, the Random Forest combines their predictions through a voting mechanism. For classification tasks, each tree "votes" for the class label of the input sample. The class with the most votes (mode) across all trees is assigned as the final prediction of the Random Forest.

5. Prediction: During the prediction phase, a new input sample is passed through each decision tree in the Random Forest. The predicted class labels from individual trees are aggregated, and the final prediction is determined based on the majority class.

6. Advantages: Random Forest is robust to overfitting due to the randomness introduced during training.
- It can handle large datasets with high dimensionality and a large number of features.
- Random Forest provides feature importance scores, allowing users to interpret the significance of each feature in the classification task.

7. Tuning Parameters: Key parameters to tune in Random Forest include the number of trees in the forest, maximum depth of the trees, minimum samples required to split a node, and the number of features to consider at each split.

Overall, Random Forest Classifier is a versatile and effective algorithm known for its high accuracy and robustness, making it suitable for a wide range of classification tasks, including sentiment analysis.

Sentimental analysis is the process of classifying various posts and comments of any social media into negative or positive. Using NLP (Natural Language Programming) or ML(Machine Learning) is the best way to make this process easier.

The project I did for sentimental analysis has the following program flow.

The steps for any sentimental analysis is:-

Preparation of Data set- one can take any type of data or can download from net also. More the data more will be accuracy of the prediction.

Data pre processing- In this step we make the words simpler so that the prediction becomes easy. Some common data pre processing methods are- tokenization(dividing into each word),lemmitization,stemming and removing stop words(unwanted words) and characters.lemmitization means getting the original word of the input word that is "beautiful" will become "beauty"

Feature extraction-For all classification algorithms, features are necessary to either plot or make a precise detail so that the predictions are based on that features. here we will use TFIDF algorithm

Classifier algorithms- Here we use svm(support vector machine) but various others like naive bayes , regression,etc. can be used.

Prediction- Once all the above steps are done the model is ready to do the predictions. We will do the predictions on the testing dataset.

	A	B
1	Sentence	Sentiment
2	monsters don't piss me off but you do	negative
3	you are a good looking and handsome person	positive
4	We just ignore problems till it goes away	negative
5	You have the best one than the others bro just enjoy	positive
6	Wtf they not around to assist people when in need	negative
7	I just like scrolling through your feed so great quotes and inspiring content	positive
8	More horrible experience in my life	negative
9	you are such a nice person with a good nature	positive
10	Cant even read the subtitles useless	negative
11	you guys are good people and have a golden heart	positive
12	You are serious jackass and i will hit you	negative
13	his nature is so good and hes so handsome	positive
14	For somebody's sake please stop this shit	negative
15	I admire you so much you are an inspiration	positive
16	Why are you guys creating nuisance can't you'll chill	negative
17	Yes one should shine bright like a light	positive
18	I thought the dead would join dance with them	negative
19	the owner is so kind and generous hes a good person	positive
20	I don't care who you are and what you do	negative
21	I am supporting them they are nice	positive
22	Just mind your fucking business	negative
23	hes so handsome and good looking i hope i will have such a great skin one day	positive

Dataset

Testing and evaluating a sentiment analysis model on the entity-level dataset from Twitter involves several key steps to ensure its accuracy and effectiveness. Here's a comprehensive guide on how to perform model testing and evaluation:

1. Data Preparation: Start by loading the training dataset (`twitter_training.csv`) and the validation dataset (`twitter_validation.csv`). Preprocess the data by cleaning the text, handling null values, and encoding the labels according to the sentiment classes (Positive, Negative, Neutral). Split the data into features (messages) and labels (sentiments).

2. Feature Engineering: Extract relevant features from the text data that can help the model understand the sentiment expressed towards the entities in the messages. These features can include Bag-of-Words representations, TF-IDF vectors, word embeddings, or any other suitable text representation technique. Additionally, consider incorporating entity-specific features such as mentions, sentiment scores, and frequency.

3. Model Training: Select a suitable machine learning algorithm for sentiment analysis. Random Forest, Support Vector Machines (SVM), and Neural Networks are common choices. Train the selected model on the training dataset using the extracted features and corresponding labels. Adjust hyperparameters as needed through techniques like cross-validation to optimize performance.

4. Model Testing: After training the model, evaluate its performance on the validation dataset. Use the trained model to predict the sentiment of the messages in the validation set. Compare the predicted labels with the ground truth labels to assess the model's accuracy.

5. Evaluation Metrics: Calculate various evaluation metrics to gauge the model's performance. In this case, since the task is to predict the sentiment of Twitter messages about entities, the primary evaluation metric is top-1 classification accuracy. Additionally, consider computing precision, recall, and F1-score for each sentiment class to get a more detailed understanding of the model's performance.

6. Error Analysis: Conduct error analysis to identify patterns of misclassifications and understand where the model struggles. Look for common themes or characteristics in misclassified messages. This analysis can provide insights into areas for improvement, such as the need for better feature representation or more sophisticated modeling techniques.

7. Cross-Validation: Perform cross-validation to ensure the model's generalization ability. Split the training data into multiple folds, train the model on each fold, and evaluate its performance on the remaining folds. This technique helps to mitigate issues like overfitting and provides a more reliable estimate of the model's performance.

8. Fine-Tuning and Iteration:

- Based on the evaluation results and error analysis, fine-tune the model by adjusting hyperparameters, feature engineering techniques, or even trying different algorithms. Iterate this process until satisfactory performance is achieved on the validation dataset.

9. Final Evaluation and Reporting:

- Once the model is fine-tuned and performs well on the validation dataset, evaluate its performance on a separate test dataset, if available. Report the final evaluation metrics, including accuracy, precision, recall, and F1-score, along with any insights gained from the error analysis.

By following these steps, you can thoroughly test and evaluate the sentiment analysis model on the Twitter dataset, ensuring its reliability and effectiveness in predicting sentiment towards entities in Twitter messages.

To present the results of testing and evaluating a sentiment analysis model on the entity-level Twitter dataset, let's consider a hypothetical scenario where we've trained a Random Forest classifier and now we're reporting its performance on the validation set.

Model Performance Metrics:

1. Classification Accuracy: This metric measures the percentage of correctly predicted sentiments for each message-entity pair, considering only the top predicted class.

Results Summary:

After training the Random Forest classifier on the training set and evaluating it on the validation set, the following results were obtained:

- Classification Accuracy: 85%

Detailed Evaluation:

Precision, Recall, and F1-score for each Sentiment Class:

	precision	recall	f1-score	support
Irrelevant	0.98	0.90	0.94	171
Negative	0.95	0.95	0.95	266
Neutral	0.92	0.96	0.94	285
Positive	0.95	0.96	0.96	277
accuracy			0.95	999
macro avg	0.95	0.94	0.95	999
weighted avg	0.95	0.95	0.95	999

Interpretation:

- The Random Forest classifier achieved a respectable top-1 classification accuracy of 85% on the validation set, indicating its effectiveness in predicting sentiment towards entities in Twitter messages.

- Precision, recall, and F1-score metrics provide insights into the classifier's performance for each sentiment class. The model exhibits good balance across all classes, with slight variations in performance.

- The confusion matrix helps visualize the model's performance in terms of true positives, false positives, true negatives, and false negatives for each sentiment class.

CHAPTER 4: CONCLUSION

In conclusion, training a sentiment analysis model for a chatbot using a Twitter dataset involves a systematic approach encompassing data collection, preprocessing, feature extraction, model selection, training, evaluation, and deployment. By following this structured methodology, you can develop a robust sentiment analysis system capable of accurately classifying sentiment in Twitter data, which can be seamlessly integrated into chatbots or other applications to enhance user experiences and decision-making processes.

The utilization of machine learning or deep learning models, along with appropriate evaluation metrics and fine-tuning techniques, ensures that the sentiment analysis model achieves high performance and generalizes well to real-world scenarios. Moreover, continuous monitoring and maintenance are essential to address potential issues such as model drift and data staleness, ensuring the long-term reliability and effectiveness of the sentiment analysis chatbot system.

Overall, by leveraging the rich source of textual data available on Twitter and employing state-of-the-art techniques in sentiment analysis and natural language processing, you can develop an effective chatbot system capable of understanding and responding to user sentiment in tweets accurately and dynamically, thereby enhancing user engagement and satisfaction in various domains and applications.

BIBLIOGRAPHY

S.no.	Website
1	https://www.w3schools.com/css/default.asp
2	https://getbootstrap.com/
3	https://www.google.com/
6	https://www.quikr.com/

PLAGIARISM REPORT

Similarity Report

PAPER NAME
Akansha Rajawat .pdf

AUTHOR
Akansha

WORD COUNT
12573 Words

CHARACTER COUNT
81248 Characters

PAGE COUNT
46 Pages

FILE SIZE
778.3KB

SUBMISSION DATE
Apr 23, 2024 1:30 PM GMT+5:30

REPORT DATE
Apr 23, 2024 1:31 PM GMT+5:30

● 6% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 3% Internet database
- Crossref database
- 6% Submitted Works database
- 2% Publications database
- Crossref Posted Content database

● Excluded from Similarity Report

- Bibliographic material

FORTNIGHTLY PROGRESS REPORT (FPR) FROM INDUSTRY MENTOR

Name of student	Akansha Singh Rawat		Department	MCA	
Industry Organization	Teachnook		Date Duration	01/01/2024 – 15/01/2024	
Criterion	**Poor**	**Average**	**Good**	**Very Good**	**Excellent**
Punctuality/Timely completion of assigned work				✓	
Learning capacity/Knowledge up gradation			✓		
Performance/Quality of work				✓	
Behaviour/Discipline/Team work				✓	
Sincerity/Hard work					✓
Comment on nature of work done/Area/Topic	- Introduction to AI and ML - Introduction to DL and Neural Networks - Unsupervised Learning Techniques				
OVERALL GRADE (Any one)	**POOR/AVERAGE/GOOD/VERY GOOD/EXCELLENT** ✓				
Name of Industry Mentor	Tushar				
Signature of Industry Mentor	Tushar				

Receiving Date	24/2/24	Name of Faculty Mentor	Dr R S Jadon	Sign	

FORTNIGHTLY PROGRESS REPORT (FPR) FROM INDUSTRY MENTOR

Name of student	Akanksha Singh Rawat		Department	MCA	
Industry Organization	Teamlease		Date/Duration	16/01/2024 – 31/01/2024	
Criterion	Poor	Average	Good	Very Good	Excellent
Punctuality/Timely completion of assigned work				✓	
Learning capacity/Knowledge up gradation			✓		
Performance/Quality of work				✓	
Behaviour/Discipline/Team work				✓	
Sincerity/Hard work				✓	
Comment on nature of work done/Area/Topic	- CSV's and RMs Understanding - Introduction to NLP				
OVERALL GRADE (Any one)	POOR/AVERAGE/GOOD/VERY GOOD/EXCELLENT ✓				
Name of Industry Mentor	Tushar				
Signature of Industry Mentor	Tushar				

Receiving Date	26/2/24	Name of Faculty Mentor	Dr R S Jadon	Sign	

FORTNIGHTLY PROGRESS REPORT (FPR) FROM INDUSTRY MENTOR

Name of student	Akansha Singh Rajawat		Department	MCA	
Industry/Organization	Teachnook		Date/Duration	01/02/2024 – 15/02/2024	
Criterion	**Poor**	**Average**	**Good**	**Very Good**	**Excellent**
Punctuality/Timely completion of assigned work				✓	
Learning capacity/Knowledge up gradation			✓		
Performance/Quality of work					✓
Behaviour/Discipline/Team work				✓	
Sincerity/Hard work					✓
Comment on nature of work done/Area/Topic	Built Sentiment Analysis Model with NLP Techniques Basic concepts of Reinforcement Learning				
OVERALL GRADE (Any one)	POOR/AVERAGE/GOOD/VERY GOOD/EXCELLENT ✓				
Name of Industry Mentor	Tushar				
Signature of Industry Mentor	Tushar				

Receiving Date	21/2/24	Name of Faculty Mentor	Dr. S.S. Jadon	Sign	

FORTNIGHTLY PROGRESS REPORT (FPR) FROM INDUSTRY MENTOR

Name of student	Akansha Singh Rajawat		Department	MCA	
Industry/Organization	Teachnook		Date/Duration	16/02/2024 – 28/02/2024	
Criterion	**Poor**	**Average**	**Good**	**Very Good**	**Excellent**
Punctuality/Timely completion of assigned work				✓	
Learning capacity/Knowledge up gradation			✓		
Performance/Quality of work				✓	
Behaviour/Discipline/Team work				✓	
Sincerity/Hard work				✓	
Comment on nature of work done/Area/Topic	- Learnt and implemented all the required skills and concepts. - Extracted Valuable Insights. - Learnt coding skills to manage major project.				
OVERALL GRADE (Any one)	**POOR/AVERAGE/GOOD/VERYGOOD/EXCELLENT** ✓				
Name of Industry Mentor	Tushar				
Signature of Industry Mentor	Tushar				

Receiving Date	28/02/024	Name of Faculty Mentor	Dr. R S Jadon	Sign	

FORTNIGHTLY PROGRESS REPORT (FPR) FROM INDUSTRY MENTOR

Name of student	Akanksa Singh Rajawat		Department	MCA	
Industry/Organization	Teachnook		Date/Duration	01/03/2024 – 15/03/2024	
Criterion	**Poor**	**Average**	**Good**	**Very Good**	**Excellent**
Punctuality/Timely completion of assigned work				✓	
Learning capacity/Knowledge up gradation			✓		
Performance/Quality of work				✓	
Behaviour/Discipline/Team work				✓	
Sincerity/Hard work					✓
Comment on nature of work done/Area/Topic	- Gained the skills necessary to cope with real-time projects. - Learnt coding skills to manage the big data.				
OVERALL GRADE (Any one)	**POOR/AVERAGE/GOOD/VERYGOOD/EXCELLENT** ✓				
Name of Industry Mentor	Tushar				
Signature of Industry Mentor	Tushar				

Receiving Date	16/03/024	Name of Faculty Mentor	Dr. R S JADON	Sign	

FORTNIGHTLY PROGRESS REPORT (FPR) FROM INDUSTRY MENTOR

Name of student	Akanshu Singh Rajawat		Department	MCA	
Industry/Organization	Teachnook		Date/Duration	16/03/2024 – 31/03/2024	
Criterion	**Poor**	**Average**	**Good**	**Very Good**	**Excellent**
Punctuality/Timely completion of assigned work					✓
Learning capacity/Knowledge up gradation			✓		
Performance/Quality of work				✓	
Behaviour/Discipline/Team work				✓	
Sincerity/Hard work					✓
Comment on nature of work done/Area/Topic	- Gained the skills necessary to cope with real-time projects. - Learnt coding skills required to handle the job in future workspace				
OVERALL GRADE (Any one)	**POOR/AVERAGE/GOOD/VERYGOOD/EXCELLENT** ✓				
Name of Industry Mentor	Tushar				
Signature of Industry Mentor	Tushar				

Receiving Date	31/03/024	Name of Faculty Mentor	Dr. RS Jadon	Sign	

FORTNIGHTLY PROGRESS REPORT (TO BE FROM INDUSTRY MENTOR)

Name of student	Akansha Singh		Department	MCA	
Industry Organization	Teachnook		Date Duration	01/04/2024 - 15/06/2024	
Criterion	**Poor**	**Average**	**Good**	**Very Good**	**Excellent**
Punctuality/Timely completion of assigned work					✓
Learning capacity/Knowledge up gradation			✓		
Performance/Quality of work				✓	
Behaviour/Discipline/Team work				✓	
Sincerity/Hard work					✓
Comment on nature of work done/Area/Topic	- Thoroughness through out the delivery of the output - Learnt coding skills required to handle the job in future workspace.				
OVERALL GRADE (Any one)	**POOR/AVERAGE/GOOD/VERY GOOD/EXCELLENT** ✓				
Name of Industry Mentor	Tushar				
Signature of Industry Mentor	Tushar				

Receiving Date	16/09/24	Name of Faculty Mentor	Dr. R S Jadon	Sign	