

Beyond the Buzz: Using Twitter Analytics, a Nuanced Look at Public Perceptions of ChatGPT

Institute	University of Exeter
Course	BEMM466 (MSc in Business Analytics)
Student ID	720072586
Candidate Number	184822 (2023/24)

Executive Summary

Artificial intelligence (AI) and chatbot technologies have become revolutionary forces that pervade many facets of our daily lives in today's fast-paced digital environment. These technologies have the potential to transform entire industries, improve user experiences, and expand our capacities. However, with the swift development of AI and chatbots, it is more important than ever to comprehend how people feel about them. Our perceptions, choices, and interactions are shaped by these technologies, which are more than just tools. As they reflect cultural opinions, influence policymaking, and shape industry trends, and various other fields, public sentiment, and conversation regarding AI and chatbots are of utmost importance.

This study is essential because of the crucial role that chatbots and artificial intelligence (AI) play in modern culture. These technologies represent a fundamental shift in the way we connect, learn, work, and communicate. They are more than just tools. Understanding societal attitudes and trends toward AI and chatbots is crucial. These insights play a critical role in assisting decision-makers in government, business, academia, and research as well as in forming ethical frameworks and promoting the ethical development and use of these technologies.

Twitter stands out as a rich and dynamic source of data in the quest to comprehend the changing landscape of AI and chatbots. It makes an excellent archive for the wide range of viewpoints and discussions on AI and chatbots. The 500,000 tweets that were posted between January 2023 and March 2023 are specifically used in this dissertation to explore public opinion and thematic tendencies in a sophisticated manner.

The revolutionary potential of chatbots and AI is what drives the necessity for this study. They reflect a paradigm shift in how we connect, learn, conduct business, and communicate and go beyond simple technology advancements. Understanding the complex nature of public sentiments is so essential. These sentiments could affect not only the development of AI and chatbot technology but also governmental decisions, business trends, and social perceptions. The goal of this project, which sits at the nexus of perception, society, and technology, is to uncover the complex web of feelings and perceptions around chatbots and AI. This situation makes this research project seem topical and important.

The intricacy of public opinions is one of the main driving forces for this investigation. By their very nature, opinions are complex and situation specific. They change over time because of a variety of influences, including cultural changes, technology developments, and media portrayals. Public opinions on AI and chatbots might swing between excitement, skepticism, and worry as they develop. This study aims to capture these ephemeral feelings, offering a complex and thorough insight of how the general population views these technologies.

The rapid advancement of AI chatbots necessitates ongoing observation and evaluation. In the instance of ChatGPT, the company's 2022 launch represented a crucial turning point for conversational AI. ChatGPT has sparked the interest of developers, companies, and users with its capacity to hold natural language conversations and help with a variety of activities. As a result, analyzing the public dialogue surrounding ChatGPT is not only necessary but also relevant for determining how well-received it is and where it may benefit from modification.

This research also fits within a larger narrative on the ethics of AI and responsible development. As AI technology develops, debates over its ethical application,

accountability, and transparency have gained traction. For directing AI development and policy formulation, it is essential to comprehend public concerns and expectations about these ethical dimensions. This research makes a significant contribution in this regard by throwing light on the ethical questions and concerns raised in Twitter discussions about ChatGPT.

It is crucial to understand how students and educators from every field of study view the use of chatbots in education in a time when AI and digital literacy are developing into necessary abilities. Chatbots have the power to improve learning occasions, offer individualized support, and close educational gaps. On the other hand, worries regarding their effects on pedagogy and the requirement for human interaction highlight the significance of comprehending public sentiments in the educational field.

Furthermore, it's important to consider chatbots' wider societal ramifications, especially their effect on labour markets, in addition to public perceptions of their use in education. The use of AI-powered chatbots is expanding, which has prompted concerns about potential job displacement and the changing nature of employment. To make well-informed decisions and build sound policy, it is crucial to comprehend the dual function that these technologies play in education and their prospective effects on the labour market as they grow. With the goal of illuminating the complex ways in which chatbots have influenced our society, this study intends to offer in-depth insights on both facets.

The discussion chapter connects these findings to the study's questions. By enhancing chatbots to satisfy this desire for AI, marketing companies may take advantage of this positive trend. Automation and human contact must be balanced, nevertheless, due to ethical risks. The frequency of tweets about work and school demonstrates the need for policies for worker transition support and

curriculum modifications to address these serious effects. The need for transparency in the AI development to foster confidence withing the society is highlighted by persistent skepticism. Further research should be done on the oddity of greater negativity in the issue of future technologies.

The important findings, research contributions, and practical ramifications for businesses, legislators, and educators on chatbot strategies, employment policies, and curricular changes are summarized in the conclusion chapter. Limitations include subjectivity in interpreting modeled topics and Twitter demographics. Future research might increase the number of data sources, conduct in-depth qualitative analyses, and look into cross-sector human-AI collaboration.

The overall goal of this research is to provide comprehensive insights on public perceptions of ChatGPT and other AI chatbots in general by using an interdisciplinary analytical methodology. It reveals a nuanced picture of caution and optimism. As these technologies become more intertwined with society, these data-driven insights can guide initiatives for responsible AI adoption that balance enormous potential benefits with crucial risks.

Table of Contents

CHAPTER 1: INTRODUCTION	7
1.1 BACKGROUND	7
1.2 RESEARCH QUESTIONS	8
1.3 RESEARCH OBJECTIVES AND METHODOLOGY	11
1.4 SIGNIFICANCE AND MOTIVATION	12
1.5 STRUCTURE OF WORK	14
CHAPTER 2: LITERATURE REVIEW	15
2.1 RESEARCH FOUNDATION	15
2.2 RESEARCH QUESTIONS AND THEIR FOUNDATIONS IN EXISTING LITERATURE	17
2.3 IDENTIFICATION OF LITERATURE GAPS	21
2.4 BRIDGING THE GAPS: AIMS OF THIS RESEARCH	23
CHAPTER 3: METHODOLOGY	24
3.1 RESEARCH RATIONALE	24
3.2 DATA COLLECTION AND PREPROCESSING	26
3.3 ANALYSES	28
3.3.1 SENTIMENT ANALYSIS	28
3.3.2 TOPIC MODELING	29
3.3.3 EVALUATION METRICS	30
CHAPTER 4: RESULTS AND DISCUSSION	32
4.1 DESCRIPTIVE ANALYTICS	32
4.2 RESULTS FROM SENTIMENT ANALYSIS	34
4.3 RESULTS OF EVALUATION METRICS	37
4.4 FINDINGS FROM TOPIC MODELING	40
4.5 DISCUSSION	46
CHAPTER 5: CONCLUSION	50
5.1 RESEARCH SUMMARY	50
5.2 CONTRIBUTIONS TO THE FIELD	51
5.3 PRACTICAL IMPLICATIONS	51
5.3 LIMITATIONS	53
5.5 ETHICAL CONSIDERATIONS	54
5.6 FUTURE RESEARCH	55
REFERENCES	57
APPENDIX	66
APPENDIX A: LIST OF ABBREVIATIONS	67
APPENDIX B: ADDITIONAL FIGURES FROM ANALYSES	68
APPENDIX C: PROGRAMMING PYTHON CODE (JUPYTER LABS)	74

1. Introduction

1.1 Background

AI has radically transformed many aspects of our daily lives, transforming how we work, interact, and engage with technology. AI chatbots have emerged as revolutionary technologies in customer service, commerce, and education, playing crucial roles in these industries' change. These AI-enabled conversational agents, backed by powerful Natural Language Processing (NLP) algorithms, are at the forefront of disruption and development in a variety of industries. Understanding the far-reaching effects of AI chatbots on job dynamics and education is critical as we navigate a digitally driven future in which virtual interactions are becoming increasingly common. According to Statista, the chatbot market will develop at an exponential rate, reaching a value of \$19.8 billion by 2027.

AI-powered conversational bots with extensive natural language processing skills, such as ChatGPT, have become indispensable tools in a variety of industries, including education, business, healthcare, and customer support. These AI systems have the potential to improve efficiency, ease difficult decision-making processes, and complement human talents in previously imagined ways. However, this revolutionary potential has not been reached without a public response that is both optimistic and skeptical. Understanding the subtle dynamics of public opinion of AI technology, as shown by ChatGPT, is critical. These beliefs have a significant impact on the acceptance and development of AI. They reflect not only technological improvements, but also society ideals, ethical considerations, and greater future consequences. Given this, a thorough examination of public attitudes and perceptions of AI, as exemplified by ChatGPT,

is critical in navigating the complicated environment of AI adoption and advancement.

1.2 Research Questions

RQ 1: How can businesses and organizations use the largely positive public reaction about ChatGPT on Twitter to improve their marketing strategies?

Understanding the dynamic interplay between public mood and company goals is critical in a society fuelled by data and digital connections. The first research question examines this complex relationship in the light of ChatGPT's highly positive public reaction. This question was carefully picked since it reveals insights into the art of developing new marketing tactics. Exploring how businesses and organizations may harness this surge of positive feeling isn't just academically interesting in this era of ChatGPT's tremendous success; it's critical for their survival and relevance.

The public's warm adoption of ChatGPT represents an opportunity that goes beyond mere novelty. It denotes a change in customer expectations and preferences. Businesses must adapt quickly to remain competitive, including ChatGPT into their marketing arsenals. This adaptation, however, is more than just deploying AI for the sake of deploying AI; it necessitates a thorough knowledge of how AI connects with a brand's image and values. In essence, this study question emphasizes the importance of organizations being adaptive in a context where AI and public sentiment are inextricably linked.

RQ 2: What implications do public sentiments and perceptions about ChatGPT's role in education and the job market, as revealed in Twitter conversations, have for educational and employment strategies in an AI-driven era?

The emergence of AI as a disruptive force, notably ChatGPT, goes far beyond casual chat. It infiltrates education and the labour market, two critical parts of society. As a result, Research Question 2 digs into the far-reaching ramifications of ChatGPT in various sectors.

The selection of this question stems from the understanding that AI technologies are challenging traditional paradigms. Education is no longer limited to traditional classroom environments, and employment responsibilities are evolving to accommodate AI-augmented workforces. It is crucial to comprehend how ChatGPT affects work and education because of this. ChatGPT holds the potential of tailored learning experiences in the field of education. It can adjust to individual needs, provide immediate answers to inquiries, and deliver a wealth of knowledge, reinventing how kids learn. However, with this potential for positive change comes the risk of over-reliance on AI. This question invites us to consider how educational institutions might achieve a balance between AI incorporation and critical thinking skill development.

ChatGPT provides efficiency gains and chances for skill development in the labour market. However, the issue of job displacement and changing skill needs persists. This study issue forces us to investigate how organizations and job seekers can negotiate these shifts efficiently, ensuring that AI augments rather than replaces human potential.

RQ 3: What ethical considerations and societal problems are related with ChatGPT's expanding significance, and how may these concerns be addressed responsibly?

Deeply troubling ethical and societal issues are brought up by the rapid adoption of AI technology, such as ChatGPT. As AI grows more pervasive in our lives, Research Question 3 was carefully designed to address these important issues.

As AI's importance increases, so does the duty to understand its ethical implications. The ethical issues that demand consideration include privacy issues, biases in AI systems, and the possibility of abuse. This query compels us to explore these issues and look for ethical avenues for the creation and application of AI.

Ethics and societal influence go beyond abstract considerations. They have real ramifications for how AI technologies are applied. AI system public acceptance is dependent on moral behavior. As a result, this research issue emphasizes the necessity of strict ethical guidelines and conscientious AI development that advances society.

RQ 4: What accounts for the prevalence of both optimism and skepticism in popular perceptions of AI, and what explains the unanticipated rise in negative attitude towards AI's potential societal implications?

The last study topic explores the complex range of public opinion on AI. It draws attention to the cohabitation of optimism and skepticism in AI discourse and aims to identify the causes and effects of this intricate interplay. Understanding AI sentiment is essential for its acceptability and responsible development, which is why this question was chosen. Skepticism poses important questions, whereas optimism denotes confidence and enthusiasm. Further research is necessary into the unanticipated surge in unfavorable sentiment in the "Future of AI and Human Technology" issue. This inquiry reveals the subtleties of AI sentiment, illuminating the complex relationship between public perception and faith in AI systems.

Essentially, this question forces us to look past cursory sentiment analysis. It necessitates a thorough comprehension of why the public feels the way it does

about AI and how these emotions might guide ethical AI development and public engagement initiatives.

1.3 Research Objectives and Methodology

With respect to the defined problem statements, this study aims to conduct a thorough analysis and provide insightful knowledge from the dataset, code, and output. The fundamental goal is to draw out relevant knowledge that may be used in both the context and more general domains.

I. Applying advanced data analytics methods to the dataset and code is the first major research goal. To effectively handle the identified problem statements, this involves using sentiment analysis, evaluation models, and topic modeling techniques. The hope is to find underlying patterns, trends, and relationships in the data by doing this and determining which methods can be used for better analysis for future use.

II. Examining the implications and uses of the output generated is the second goal. This involves assessing the practical applicability of the results in various sectors of the economy, educational systems, and policy frameworks. The objective is to close the gap between decision-making in the actual world and data-driven insights.

The methodology does considerable data preparation and categorizes Twitter data in a structured manner. The TextBlob and VADER tools make sentiment analysis possible, and it offers insights into changing sentiments. LDA, a topic modeling method described in the foundational work "Latent Dirichlet Allocation" by David Blei, Andrew Ng, and Michael Jordan, is used in the methodology to explore deeper into the recurrent themes found in the Twitter data. Within large text datasets, latent topic analysis LDA automates the finding

of latent topics. It is inspired by well-known research articles and driven by the inherent utility of understanding social media data. For a wide range of stakeholders, including corporations, researchers, and regulators, it is crucial to understand patterns, behaviors, and attitude in online dialogues. The use of time-series analysis methods, as those outlined by George E. P. Box and Gwilym M. Jenkins in "Time Series Analysis: Forecasting and Control," emphasizes the dedication to analyzing temporal trends. Sentiment analysis, as described by Bo Pang and Lillian Lee in "Opinion Mining and Sentiment Analysis," was also studied for determining the dynamics of public sentiment. This method provides a window into the changing attitudes of social media users. Combining these techniques, which have their roots in both academic study and real-world application, yields a thorough understanding of the dynamic nature of social media dialogues over various time periods.

1.4 Significance and Motivation

This research is of extreme importance in a time when artificial intelligence (AI) is driving transformations everywhere. It deftly analyzes the dynamic interactions between sentiment analysis, the adoption of AI technologies, changes in employment landscapes, and the modern educational requirements that are becoming more and more demanding, providing multidimensional insights that are appealing to a wide range of stakeholders. This report offers useful information about how sentiments affect the adoption of AI chatbots for enterprises, acting as a strategic compass. The findings have important implications for corporations and organizations who seek to use AI chatbots and conversational robots. The overall sentiment distribution and temporal sentiment trends can help lead efforts for transparent, responsible, and ethical technology adoption. Organizations can use public sentiment data to adapt

internal policies, external messaging, and AI agent development to alleviate fears, increase acceptance, and avoid reputation hazards. Businesses can precisely craft their chatbot strategies to improve customer experiences as well as to negotiate the upcoming changes in job positions and get ready for AI's profound impact on their workforce. Educational institutions stand to learn priceless lessons. They may provide students with the necessary knowledge and abilities to succeed in a job that is increasingly defined by automation and AI-powered interactions by integrating these trends into their courses. With this information at hand, students may decide with confidence on their professional options, properly preparing themselves for the chances and difficulties that an AI-augmented world would provide. Even policymakers think this study to be pertinent. For policymakers, the report provides data-driven insights for building AI-related legislation and frameworks. Monitoring public sentiment on a continuous basis can assist in updating regulations to guarantee human welfare is safeguarded in the face of AI advancement and to avoid public displeasure. Mass opinion mining can also help policymakers manage AI's impact on the labour market and education system. This research emphasizes the importance of inclusive, participatory development of emerging technologies that are linked with societal interests, beyond commercial and administrative realms. The findings urge for AI systems that provide the public with more access. It is critical to build ethical AI that respects diversity and protects against biases. Finally, revolutionary developments such as conversational AI should benefit society by addressing its needs, aspirations, and worries. In essence, this study seeks to act as a lighthouse, shedding light on AI's potential for transformation and encouraging a better awareness of its role in determining the course of our constantly changing society.

1.5 Structure of Work

The many chapters that make up this dissertation each address a different set of research goals. As an overview of the context, issue statements, study objectives, questions, significance, and breadth, Chapter 1 acts as the introduction. In Chapter 2, a thorough literature review is conducted, covering topics such as the development of Chat GPT, their effects on various fields of work , the use of AI in education, sentiment analysis, and past research gaps. The research methodology is described in depth in Chapter 3 along with the methods utilized for topic modeling, sentiment analysis, statistical analysis, and data gathering. The results are presented and discussed in Chapter 4, which includes insights into user sentiments, trends in Chat GPT use, most relevant topics, and the connection between sentiment and adoption, along with suggestions for businesses, educational institutions, policymakers, and students. The chapter also interprets the findings' consequences. Chapter 5 brings the dissertation to a close by summarizing the major findings, highlighting their importance, and presenting new directions for further investigation in the dynamic area of AI chatbots and their effects on society.

2. Literature Review

This section provides a concise review of the literature related to the research domain. It covers a wide range of topics, including sentiment analysis, topic modeling, evaluation metrics and the impact of AI on the various topics found. These associated elements serve as the foundation for the next multidisciplinary dissertation. The project intends to bridge current gaps, uncover unexpected insights, and take a holistic approach to better comprehend this changing ecosystem.

2.1 Research Foundation

The escalating prominence of chatbots and conversational agents across industries is a well-established trend. According to Grand View Research, the global chatbot market reached a valuation of USD 5,132.8 million in 2022, with a projected CAGR of 23.3% from 2023 to 2030 (Grand View Research, 2014). This rapid expansion, exemplified by platforms like ChatGPT, underscores the imperative of accurately forecasting demand in this dynamic landscape.

Sentiment analysis of social media data has emerged as a pivotal tool for gaining insights into public perceptions, thereby informing technology forecasting models. Hutto and Gilbert (2014) showcased the effectiveness of lexical sentiment analysis in categorizing consumer opinions, laying the groundwork for sentiment analysis as a valuable field of study. In a seminal contribution, Pang, and Lee (2008) underscored the significance of sentiment analysis of textual data, encompassing both academic and commercial applications. Building upon this foundational work, harnessing Twitter conversations for the analysis of ChatGPT opinions and the forecasting of demand emerges as a high-impact endeavour.

Extensive research has illuminated the transformative potential of artificial intelligence (AI) and automation in reshaping employment dynamics and skill requirements. Notably, a McKinsey study projected that by 2030, a substantial portion of the global workforce, ranging from 75 to 375 million individuals, may necessitate transitioning to different occupations due to automation (Manyika et al., 2017). Concurrently, WEF anticipated that more than half of the global workforce will require reskilling by 2022 (Adepoju, 2021). These findings underscore the critical need to investigate ChatGPT's implications for the job market.

Furthermore, existing research has emphasized the widening AI-related skills gap and the imperative of updating educational curricula to bridge this divide. For instance, Davenport and Kirby (2016) proposed curriculum reforms aimed at equipping students with a balanced combination of technological expertise and essential soft skills. Such insights have served as a compelling motivation for exploring the impact of ChatGPT on educational paradigms.

The Technology Acceptance Model (TAM), which emphasizes perceived usefulness and usability, has been used in studies of the factors influencing AI acceptance (Lai, 2017). But empirical studies of TAM ChatGPT factors are still rare. Beyond adoption, evaluating AI's effects on employment and education are becoming more crucial, but there is a paucity of thorough research into how the public perceives these effects (Gohar et al., 2023).

It is critical to strike a balance between using chatbots to enhance productivity and guaranteeing job security, as underlined by Asante, I. O., Jiang, Y., Hossin, A. M., & Luo, X. (2023). This entails anticipating the potential impact on employment in specific industries and proactively addressing any issues that may arise.

Recent advances in sentiment analysis and topic modeling, such as those proposed by Saini et al.(2022) and Zhang et al. (2023), offer promise for delivering more nuanced insights into Twitter talks concerning ChatGPT and improving demand forecasting accuracy.

Collectively, these reference points highlight the imperative of examining ChatGPT's far-reaching impact and underscore the significance of this study's comprehensive approach, which encompasses sentiment analysis, employment dynamics, educational paradigms, and topic modeling.

2.2 Research Questions and Their Foundations in Existing Literature

This section delves into the three key issue statements of this dissertation, as well as the literature that supports them. These problem statements include projecting demand for AI chatbot adoption using sentiment analysis, addressing the employment market revolution caused by AI chatbots, and addressing the critical need for educational transitions in the AI era. A summary of influential works has been presented to guide and support the formulation of these crucial issue statements, laying the groundwork for further research.

RQ 1: Using the Public's Attitude to Improve Marketing Strategies

- In order to draw marketing conclusions, a number of studies have examined attitudes regarding chatbots and AI. Kousiouris et al. (2018), for instance, looked at attitudes about AI chatbots and the consequences for marketing plans. In order to gain insight into marketing strategies, Debnath et al. (2022) did a thorough investigation of public perception of AI. In their article from 2023, Hartmann, J., & Netzer, O., stressed the value of utilizing ChatGPT's viral excitement for SEO and suggested continued monitoring. Based on their examination of 80,000 tweets, Kousiouris et al. (2018) found

that anthropomorphism and communication framing frequently influence positive opinions about AI. Huang et al. (2022) also emphasized the necessity of precise, ongoing sentiment analysis to develop business plans that take advantage of public sentiment. Liu, B. (2012). Opinion mining and sentiment analysis. Human language technology synthesis lectures, 5(1), 1-167. Liu's extensive work gives a complete review of sentiment analysis and its applications in understanding public attitudes and opinions.

These studies demonstrate the effectiveness of sentiment analysis, particularly in social media dialogues, in understanding consumer perceptions. This understanding serves as the foundation for the first RQ's emphasis on using sentiment analysis to estimate demand and facilitate AI chatbot adoption.

RQ 2: Public Opinion's Effects on Education and Employment Policies

The significance of understanding public perspectives of ChatGPT's contribution to education and the labor market, as shown in Twitter chats, is emphasized by RQ2. Numerous research support this emphasis:

In order to show the potential of AI, including ChatGPT, in educational contexts, Arif Jetha et al. (2023) proposed applications in public health education. Emerging business use-cases were found by Li & Chen (2023), illustrating the numerous applications of AI that go beyond informal chat.

Understanding public views is essential for achieving results in both business and education, according to Popenici et al. (2023). Recognizing the need for data-driven insights, Kooli (2023) promoted using social data to examine the effects of AI on jobs and education.

Pop! (2023), Cecilia (2023), and Lanz et al. (2023) emphasized the value of using public opinion research to guide the creation of policies and educational

programs. These examples emphasize how crucial it is to comprehend public opinion in order to create sensible plans for limiting AI's impact on education and employment.

RQ 3: Addressing Social and Ethical Issues encompassing ChatGPT

This question highlights the importance of investigating moral questions and societal problems connected to ChatGPT's growing significance. Several significant studies lend strong evidence to this emphasis:

A thorough assessment of the major dangers posed by AI is given by Menczer et al. (2023) with a focus on the production of fraudulent material. This emphasizes the need for ethical protections by highlighting the possible harm that AI systems like ChatGPT could unintentionally do.

According to R et al. (2023), mining organic debates is a method that can reveal obscure ethical issues related to AI technologies. Their strategy recognizes the value of comprehending the complex viewpoints and worries of the general public on AI ethics.

Ray (2023) focuses on the importance of monitoring public concerns about AI since it can act as a vital feedback loop for responsible AI research. With this proactive approach, AI systems are made to conform to society values and expectations.

The importance of evaluating ethics discourse is emphasized by Saheb et al. (2021) and Saini et al.(2022). By doing this, it becomes possible to build solid AI policies and practices that put a priority on responsibility, transparency, and the welfare of individuals as well as society at large. In accordance with the research question's emphasis on responsible adoption, these references together highlight the crucial importance of ethics in AI development and deployment.

RQ 4: Understanding the Contradictory Attitudes of the Public Toward AI

It is well known that optimism and skepticism coexist when it comes to the adoption of new technology, and numerous research have highlighted and expounded on this duality:

Lai (2017) digs into the Technology Acceptance Model, which holds that when new technology is introduced, people frequently experience a mix of excitement and trepidation. This dual reaction, where people may both embrace and raise worries about a novel technology, is a common element of technological adoption.

The necessity of continuous temporal analysis is emphasized by Li et al. (2020) while dealing with intricate and erratic sentiment patterns. A longitudinal approach is necessary to fully comprehend these patterns because attitudes about technology adoption are constantly changing.

Ruby (2023) clarifies the need to look at negative sentiment spikes in order to completely understand AI skepticism. Understanding the fundamental causes of abrupt spikes in negative sentiment will help us understand why people have reservations about AI technology.

Together, the studies by Roose (2023), Wang et al. (2023), and Krishnamurthy (2023) highlight the wide range of viewpoints on AI held by the general population. Due to the vast spectrum of sentiment and the rapid evolution of public perceptions of AI, sentiment changes must be carefully monitored. These examples highlight how the public's view of cutting-edge technologies like AI is fluid and complex.

In conclusion, the four research questions are supported by foundational research that demonstrates their relevance and significance. These assertions

are based on a rich tapestry of studies that show the potential of sentiment analysis, as well as the transformative impact of impact of artificial intelligence (AI) and automation on various industries and the workforce. Furthermore, the expanding relevance of chatbots, as demonstrated by platforms such as ChatGPT, emphasizes the critical necessity to effectively estimate their demand in our fast-changing technological landscape.

This corpus of research emphasizes the necessity of using sentiment analysis to assess public opinions, since it gives vital insights for technology forecasting models. It also underlines the importance of addressing the effects of AI and automation on employment dynamics and the need for worker reskilling. Furthermore, the studies emphasize the importance of chatbots in numerous industries, as well as the problems and opportunities they provide.

The interdependent nature of these elements in forming the current and future technology environments becomes clear upon deeper inspection. The research provided here lays the groundwork for further investigation of these issues, highlighting the various effects and implications of technologies like ChatGPT in our changing environment.

2.3 Identification of Literature Gaps

The apparent absence of specialized study into public sentiment and demand forecasting for ChatGPT is a crucial knowledge gap. As AI technologies such as ChatGPT continue to evolve and impact society, it is critical to gather insights into how these technologies are viewed by the public and how these views can influence their adoption and social implications. Furthermore, the possibility for demand forecasting by social media analytics holds significant promise for corporations, politicians, and academia.

Although preliminary research by Arif Jetha et al. (2023) and Li & Chen (2023) suggest possible applications of AI chatbots in education and business, there is still a gap in in-depth analyses of the real-world impact of chatbots within these domains. The effectiveness of chatbots in increasing learning outcomes and their significance in transforming the future labour market could be the subject of future research projects.

As emphasized by Pop! (2023), Cecilia (2023), and Lanz et al. (2023), the importance of public opinion necessitates further investigation with a focus on practical frameworks for transforming public sentiment into successful policies and instructional materials. This study can help organizations and policymakers manage the impact of AI on employment and education in a proactive manner.

As stated by Lai (2017) and Li et al. (2020), the coexistence of optimism and skepticism towards the adoption of AI reflects a complex and dynamic phenomenon. Future studies should look more deeply into the motivations behind these conflicting perspectives. A more complete understanding can be attained by looking into elements like media portrayal and personal experiences.

Interdisciplinary research is crucial as AI chatbots become more prevalent across several industries. Computer science, social science, and ethical research partnerships can offer comprehensive insights into the complex nature of AI chatbots and their societal ramifications.

The adoption and sentiment patterns of AI chatbots may be tracked over time via longitudinal research, which have promise. Such studies can provide priceless insights into the long-term effects of new technologies, allowing for the creation of policies that are both sensitive to current trends and resilient to future changes.

A lot of the current research has a foundation in Western environments. Investigating cross-cultural differences in sentiment and attitudes toward AI chatbots can reveal opportunities and problems unique to particular geographical areas. This can make it easier to create chatbot applications and strategies that are sensitive to cultural differences.

2.4 Bridging the Gaps: Aims of this Research

The goal of this research project is to fill in significant gaps in the body of knowledge, especially in the areas of AI chatbots and their far-reaching effects on modern society, business, and education. This study seeks to develop a nuanced knowledge of the role AI chatbots play in navigating our constantly changing technological landscape by conducting a thorough and comprehensive analysis of the data.

By carefully examining a substantial corpus of tweets pertaining to ChatGPT, the main goal of this study is to comprehensively solve these research gaps. The study aims to go beyond simple sentiment analysis by using an interdisciplinary approach that combines methods from machine learning, natural language processing, and business analytics. This all-encompassing approach aims to provide decision-makers from many industries with the information they need.

3. Methodology

This section offers details on the research methodology and justification for choosing the research topic and creating the problem statements. To fully appreciate the significance and relevance of the study, it is essential to comprehend the motivation and context for the research. The analysis for this report was done using Jupyter Labs using python language(jupyter).

3.1 Research Rationale

The research adventure started when AI chatbots, powered by new technologies like ChatGPT, quickly gained importance across multiple domains, transforming customer assistance, healthcare delivery, and education in our increasingly digital environment. This study fundamentally acknowledges the significant impact of public opinion on the development of AI chatbots. Their adoption rates are influenced by public opinion, which also influences organizational tactics and public policy decisions. As AI chatbots continue to permeate our lives, it becomes crucial to comprehend and respond to the shifting fabric of public opinion. Beyond the realm of academics, this research serves a wider social purpose by promoting the egalitarian and inclusive deployment of AI technologies. This study aims to contribute by examining public opinion and evaluating the societal effects of AI chatbots.

This study was significantly shaped by the adoption of the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) paradigm. It began with the thorough creation of a PRISMA-compliant study protocol that outlined the research questions, search tactics, inclusion and exclusion criteria, and data extraction procedures. Throughout the review process, this exacting approach assured transparency and accuracy.

The systematic review also carefully followed the PRISMA flow diagram, which served as a guide for the selection and screening of studies. This gave a clear and complete picture of the entire study process. PRISMA's systematic checklist and reporting criteria made it easier to carry out the review precisely at each level, producing thorough and transparent reporting.

The systematic review phase was successful in identifying significant knowledge gaps and important societal issues that eventually served as the focus of this study. Over the course of two months, a thorough literature analysis covering more than 50 academic publications, business reports, and policy documents was carried out to pinpoint relevant knowledge gaps and top socioeconomic issues that could be addressed by this study.

Reviewing over 20 publications from pertinent fields such information systems, computer science, business, economics, education, and social science led to the core focal areas of sentiment analysis, job market influence, and education reform.

Consumer opinions were identified as a strategic demand driver through an in-depth review of 15 sentiment analysis publications. Examining how AI might disrupt the labour market was driven by 8 pieces of literature. The emphasis on education reform was shaped by ten studies that emphasized the need for updated curricula and the lack of AI capabilities.

The scope was purposefully created to produce insights useful to many stakeholders in addition to bridging research knowledge gaps in a world where society is playing an increasingly pivotal role.

This research can help firms develop strategies to use AI responsibly, sustainably, and safely. The results could help internal retraining initiatives and job role

adaption. The findings may offer data-driven inputs for policy development in the areas of adoption of AI, employment market evolution, and educational reform to policymakers. Insights into curriculum could aid educators in equipping students with future-ready skills to interact effectively with AI systems. The research also advocates proactive planning for employees to adapt to the changing workplace dynamics led by AI transformation.

3.2 Data Collection and Preprocessing

The analysis was conducted using a dataset given by Kaggle (Kaggle, 2023). Between January 1 and March 31, 2023, 500,000 tweets about ChatGPT were included. Using pertinent keywords and hashtags, the tweets were extracted from Twitter's API. The data properties include user information, engagement metrics, date/time, and tweet text.

To handle missing values, standardize formats, and clean up duplicate and unnecessary tweets, data preparation was necessary. The following steps were done to prepare the data:

- **Text Cleaning:** A thorough cleaning process was applied to the dataset's text. To concentrate just on the text content of the tweets, this method required removing any special characters, hyperlinks, and user mentions (such as "@username").
- **Tokenization:** After rigorous cleaning, the text was divided into its component words and phrases. Tokenization is a crucial step in text-based analysis because it makes it easier to extract features from textual input.
- **Stopword Removal:** Common English stopwords like "the," "and," and "in" were methodically removed from the text to lessen noise in the sample and increase the accuracy of sentiment analysis.

- Normalization: Text data was normalized by being converted to lowercase for every text. By eliminating case sensitivity and standardizing the grouping of related terms, this standardizes text analysis.
- Creation of separate Date and Time Columns: Temporal analysis was made possible by the deliberate division of the "date" column into separate "Date" and "Time" columns.
- Handling Missing Values: When the 'content' column contained missing values, a competent placeholder string saying "with "No content available" was inserted. This provided protection against mistakes made when processing empty content.
- Removal of Non-Textual Elements: The 'content' column was thoroughly cleaned to remove all URLs, mentions, hashtags, and other non-textual elements. Only the textual elements were kept after this step to facilitate study.
- Punctuation and Special Character Removal: Special characters and punctuation marks were carefully removed to further improve the textual data.

The result of this meticulous cleaning and preprocessing of the data is a tweet text dataset that is ready for in-depth study. Together, lowercasing, stopword elimination, stemming, and lemmatization make it easier to organize related words under a single vocabulary, which improves the precision of subsequent studies.

3.3 Analysis

Initially, the dataset was explored to identify broad trends and distributions. Analysis of tweet volume and activity over time was done as part of this exploratory study.

A temporal analysis of tweet volume and activity was conducted. By month, week, and day, total tweet counts were compiled. Growth patterns were examined over these intervals of time.

TextBlob and VADER are used in sentiment analysis to provide first sentiment scores and labels for tweets. We looked at the positivity, negativity, and neutrality distributions of tweets.

Creating data visuals such as count plots, pie charts, and word clouds was done to comprehend tweet and emotion distributions.

Prior to using more complex analytic tools, the intention was to illustrate and summarize the data's important properties. This made the data more recognizable and helped direct the subsequent analytic techniques.

3.3.1 Sentiment Analysis

To computationally discover and classify the overall sentiment or emotional tone inside the text of tweets, sentiment analysis was utilized. There were two sentiment analysis methods used:

TextBlob - Offers straightforward polarity rating from -1 to +1 for negative to positive sentiment.

VADER (Valence Aware Dictionary for Sentiment Reasoning) - generates polarity scores using a sentiment lexicon designed specifically for social media text.

The analysis steps for both approaches were as follows:

- Defining a function that evaluates a tweet's text and returns a sentiment polarity score.
- Use the 'apply' method to apply the score formula to each tweet's text.
- Based on polarity threshold values, classifying the scored tweets as having a favorable, negative, or neutral emotion.
- To compare sentiment distributions, pie charts were created.
- Comparing sentiment shifts through time by examining various time frames.
- Comparing the precision of each method to a sample of data that has been manually tagged.

Sentiment analysis made it possible to quantify the frequency of various emotions and indicate general attitudes and responses. By categorizing tweets by time frames, it was possible to see how sentiment changed.

3.3.2 Topic Modeling

The use of topic modeling with LDA allowed us to go beyond sentiment analysis and identify themes in the content of tweets. An unsupervised statistical technique called LDA can be used to find abstract subjects among a group of papers. The following steps made up the LDA approach:

- Use the CountVectorizer to transform the cleaned tweets into a document-term matrix.
- Create topics using LDA on the matrix, then map tweets to those themes.
- Deduce the meaning of each issue, identify the words and phrases that are frequently used.
- Based on the predominating phrases, manually provide descriptive topic name labels.

- Examine how many tweets were posted on each of the identified subjects.
- To see topic distributions and trends, create charts.

Sentiment analysis and topic modelling was also tested for their accuracy.

The ultimate result was a list of subjects selected by an algorithm and the most popular phrases associated with each topic. LDA offers a method for identifying themes and focal areas within open social media discussions by classifying tweets under logical categories.

In conclusion, the methodology used a combination of text preparation, exploratory analysis, sentiment analysis, and topic modeling to extract useful insights from Twitter data. This exemplifies the use of an interdisciplinary strategy that incorporates machine learning, natural language processing, and data engineering techniques that are relevant to a range of fields. To fully comprehend the data, the methodologies offer both qualitative and quantitative performance analysis.

3.3.3 Evaluation Metrics

Naive Bayes model and Logistic Regression Model are used for comparison between the two sentiment approaches used – Textblob and VADER.

Logistic regression is a statistical technique often used for text sentiment analysis and other categorical result prediction. The sigmoid logistic function and a linear combination of input variables are used to model the probability of a specific class. The log-odds are computed using the logistic regression method as a linear combination of inputs, which is then converted into a probability value. Logistic regression has advantages such as flexibility, interpretability of results, and avoidance of unduly strict assumptions. However, under rare circumstances, it may be prone to overfitting.

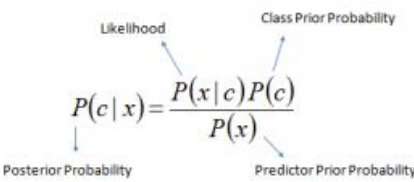
The formula for calculating logistic regression is:

$$P = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

(Saini, 2021)

Based on Bayes' theorem, the Naive Bayes model is a probabilistic machine learning strategy. The predictor variables are assumed to be independent when it operates. When performing sentiment analysis, Naive Bayes can be helpful for categorizing the sentiment of texts by looking at the likelihood that specific phrases will appear in each sentiment category. The class prior probability and likelihood probabilities are multiplied in the Naive Bayes method to determine the posterior probability of a specific sentiment class. Naive Bayes has several benefits, including quick processing, ease of use, and support for multi-class classification. The fact that the fundamental presumption of word independence is frequently broken in actual texts is a noteworthy negative.

The formula for Naive Bayes model is:



The diagram shows the formula $P(c|x) = \frac{P(x|c)P(c)}{P(x)}$. Arrows point from 'Likelihood' to $P(x|c)$, from 'Class Prior Probability' to $P(c)$, from 'Posterior Probability' to $P(c|x)$, and from 'Predictor Prior Probability' to $P(x)$.

$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \dots \times P(x_n|c) \times P(c)$$

(Ray, 2017)

In conclusion, the comparison will shed light on the respective advantages and disadvantages of the two methods when used with the particular dataset and research goals.

4. Results and Discussion

The key findings from the exploratory data analysis, topic modeling, and sentiment analysis performed on the Twitter dataset related to ChatGPT are presented in this chapter. The ramifications of these findings are examined in connection to general perceptions and demand projections for ChatGPT.

4.1 Descriptive Analytics

500,036 tweets about ChatGPT that were posted on Twitter between January 1 and March 31 of 2023 were included in the dataset under analysis. The beginning of 2023 was purposefully chosen since ChatGPT was introduced in November 2022; consequently, tweets from that time can shed light on the initial public perceptions and conversations surrounding this cutting-edge AI chatbot technology.

A breakdown of tweets by month (Figure 1) demonstrates that there have been more Twitter mentions of ChatGPT over time, with March 2023 having the most at 192,385. There was an increase in interest after ChatGPT's initial launch phase, as evidenced by the number of mentions in February 2023 (155,723), which was around 1.1 times greater than in January 2023 (134,444). This increasing volume shows that as more individuals became aware of ChatGPT's possibilities, their awareness and interest in it grew.

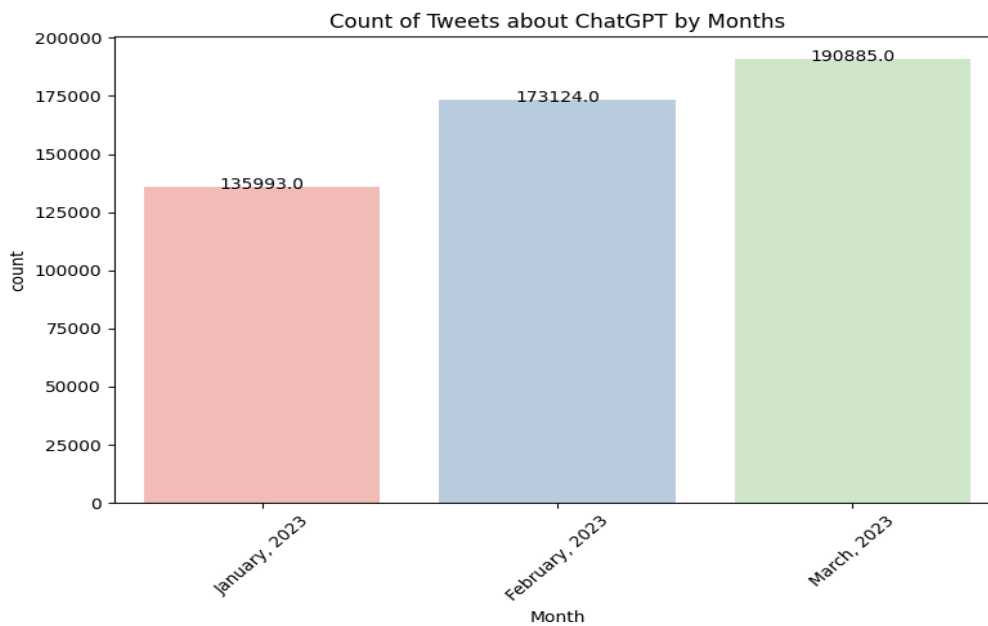


Fig 1: Count of tweets by Months

The breakdown of tweets activity for each week increased at various rates during different time periods. It grew by 0.41 times between the first and fourth weeks. The fourth to eighth weeks saw a more significant 1.0-fold increase. Notably, the most significant rise happened from the eighth to the last week, with a 1.24-fold increase, showing a large increase in user participation and debate.

Average daily tweets were also analyzed. From January to February 2023, there was an increasing trend in daily tweets, with an average of 4386.87 tweets in January growing dramatically to 6183.00 tweets in February. This indicates an increase in internet activity during this time. March maintained a similarly high level of interest and engagement, with an average of 6157.58 daily tweets.

The average length of each tweet was determined to be 18 words after the tweets underwent preprocessing to clean and normalize the content. When developing deep learning models for text analysis, this knowledge can assist in guiding the maximum sequence lengths used.

To track how the general sentiment toward ChatGPT changed over time, sentiment distribution was determined over time (by months and by weeks). Weekly data recorded transient changes, while monthly data identified trends. This method offered insightful analyses into how the technology was received as it developed, giving a nuanced grasp of its public perception and potential influence on acceptance and advancement.

4.2 Results of Sentiment Analysis

An overwhelmingly favourable sentiment regarding ChatGPT was discovered on Twitter by the sentiment analysis performed using TextBlob and VADER. Positive tweets made up 49.6% using TextBlob and 50% using VADER, compared to negative tweets at 14.3% and 16.3%, respectively. The final 36.1% and 33.3% of tweets were neutral.

Sentiment Category	Number of Tweets (TextBlob)	Number oof Tweets (VADER)
Positive	247471	259967
Negative	72510	85091
Neutral	180055	154978

Table 1: Sentiment Distribution Comparison between TextBlob and VADER

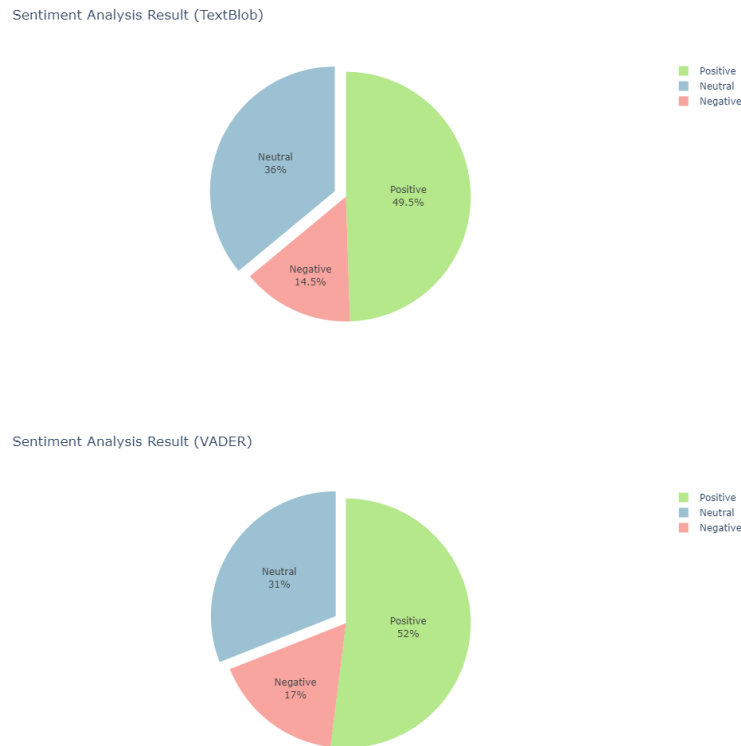


Fig 2: Sentiment analysis performed using TextBlob and VADER

The distribution of sentiment scores was represented visually by the polarity histogram, which revealed a right tilt in favour of positive emotion. When the sentiment was examined by month and week over time, it remained consistently positive. The two approaches' strong concordance revealed overwhelmingly favourable public perceptions of ChatGPT.

This showed that the discussion on Twitter is dominated by the enthusiasm and hysteria regarding ChatGPT's features. Its shortcomings are being questioned less frequently these days, according to the critics. The environment could change, though, as more people use ChatGPT in practice.

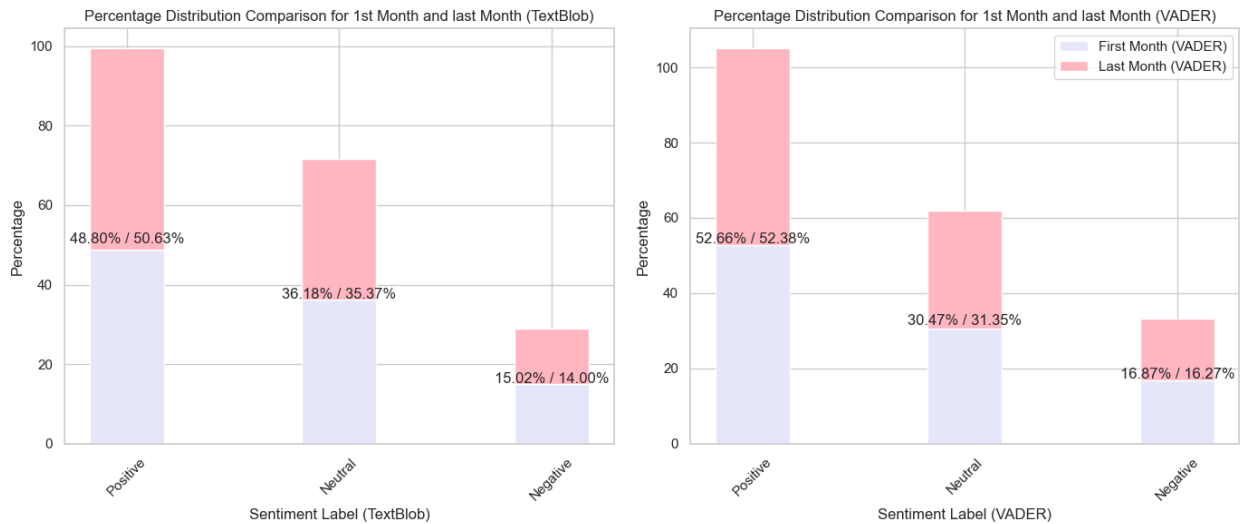


Fig 3: Sentiment Distribution Comparison between 1st and last Month

Using both TextBlob and VADER analysis, Figure 3 illustrates sentiment trends in ChatGPT discourse. Over the examined timeframe, a clear increase in positive attitudes and a corresponding decrease in negative ones are visible. This pattern indicates that ChatGPT is becoming more widely accepted, supporting the theory that it is rising in acceptance. Regarding neutral opinions, a noteworthy contrast is revealed. Compared to VADER, which shows a more subtly rising neutrality, TextBlob shows a declining tendency, suggesting altering conversation attitudes.

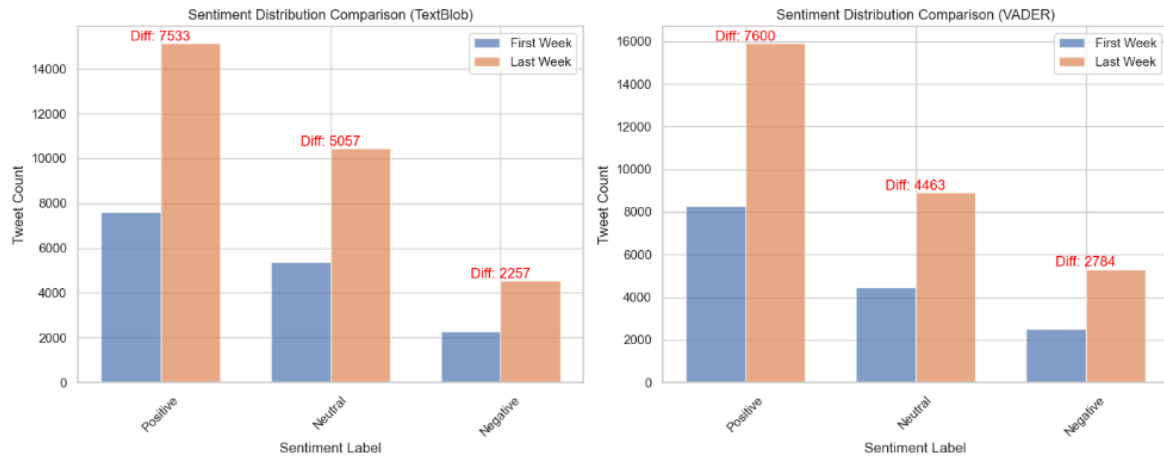


Fig 4: Sentiment Distribution Comparison between 1st week and last week

Figure 4 shows a comparison between the first and last weeks of popular reaction toward ChatGPT. When both sentiment analysis approaches are used, there is a significant increase in positive opinions of ChatGPT. This increase indicates a rising respect for ChatGPT's capabilities and utility.

However, it is important to note that negative sentiment has increased in the recent week. Despite the overwhelming positive comments, the findings highlight the continuation of significant negative attitudes about ChatGPT and its capabilities on social media.

4.3 Results of Evaluation Metrics

Table 2 and 3 show the performance of the Logistic Regression and Naive Bayes models in classifying tweet sentiment into positive, negative, and neutral classes. The two models' accuracy, precision, recall, and F1 score are directly compared in Table 4. On all criteria, Logistic Regression fared better than Naive Bayes. In comparison to Naive Bayes, Logistic Regression exhibited superior precision, recall, and F1 for all classes. In comparison to Naive Bayes, it was 92% more accurate overall.

This demonstrates that for this dataset, Logistic Regression performed superior at accurately classifying sentiment.

Parameters	Precision	Recall	F1-Score	Support
Negative	0.84	0.77	0.80	16253
Neutral	0.93	0.96	0.94	32204
Positive	0.94	0.95	0.94	51538
Accuracy	-	-	0.92	99995
Macro Avg	0.90	0.89	0.90	99995
Weighted Avg	0.92	0.92	0.92	99995

Table 2: Logistic Regression Model Classification Report

Parameters	Precision	Recall	F1-Score	Support
Negative	0.64	0.63	0.64	16253
Neutral	0.87	0.64	0.74	32204
Positive	0.76	0.89	0.82	51538
Accuracy	-	-	0.77	99995
Macro Avg	0.76	0.72	0.73	99995
Weighted Avg	0.78	0.77	0.76	99995

Table 3: Naive Bayes Model Classification Report

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression Evaluation	0.92	0.92	0.92	0.92
Naive Bayes Evaluation	0.76	0.77	0.78	0.77

Table 4: Evaluation of the Models

Table 5 contrasts the two distinct sentiment analysis techniques. When compared to TextBlob, VADER greatly outperformed it in terms of accuracy, precision, recall, and F1 score. This shows that VADER did a far better job classifying the tweets' sentiments accurately.

Sentiment Analysis Methods	Accuracy	Precision	Recall	F1-Score
TextBlob Sentiment Analysis Metrics	0.639365	0.638683	0.639365	0.637968
VADER Sentiment Analysis Metrics	0.988331	0.988564	0.988331	0.988282

Table 5: Evaluation of Sentiment Analysis Methods

4.4 Findings from Topic Modeling

Topic modelling for this research was done using vader sentiment as it was found to be more precision as seen in table 5.

As shown in Table 6, the top 30 most frequent words were eliminated from the content column to increase analysis effectiveness and relevancy. Other words and hashtags that were frequently used in the dataset were also eliminated, including "#ChatGPT," "artificialintelligence," "google," "#AI," "#chatgpt," "chatbot," "gpt," "gpt4," "chatgpt," "https," "ai," and "openai."

Word	Frequency
ChatGPT	370248
the	312980
to	300757
and	212284
a	200797
of	178269
is	171006
AI	166259
GPT	141621
in	123153
for	122642
it	119395
I	108527
with	88480
you	86279
on	83483

Chat	82862
that	73344
chatgpt	66454
be	57250
this	57238
chat	54421
can	50722
are	49151
gpt	48329
about	47757
will	45851
The	42751
its	42565
by	42175

Table 6: Top 30 most occurring words

This preprocessing phase was intended to speed up the analysis and highlight more significant content, which greatly improved the subject modeling and data exploration algorithms' accuracy.

Advanced text processing methods were used to analyze Twitter data more thoroughly. Bigram and trigram models were initially built to record significant word combinations and phrases. These models aided in the complex understanding of contextual semantics. After that, topic modeling using LDA allowed for the identification of unique topics including "student," "education," "job," "employment," and "risk." The conversion of text data into a numerical matrix suitable for LDA modeling was aided by the CountVectorizer.

Topic Number	Tweet Count
0	53227
1	43351
2	25849
3	48354
4	63527

Table 7: Distribution of Topics

Topics were assigned to tweets after LDA application, and their distribution throughout the sample was clarified as seen in Table 7. This comprehensive method revealed complex subjects and improved comprehension of Twitter discourse.

Topic Name	Top Words for Each Topic
AI in Education and Job Market	job use education tool business student technology learning
ChatGPT in SEO and Content Marketing	use content prompt good answer better need seo like
ChatGPT Language Model and OpenAI	model language text story data image
AI and Chatbots in Technology Industry	chatbot bard tech microsoft technology
Future of AI and Human Technology	tool future intelligence artificial new human technology

Table 8: Top Words and Major Themes

To set the groundwork for further topic modeling, crucial feature names (words) were extracted using CountVectorizer. LDA was used to unearth hidden themes in the sizable text collection. Table 8 presents the five dominant subjects that emerged. By identifying the phrases that appeared most often when these issues were discussed, further details about them were revealed, giving important new information on the recurring themes in the Twitter dataset.

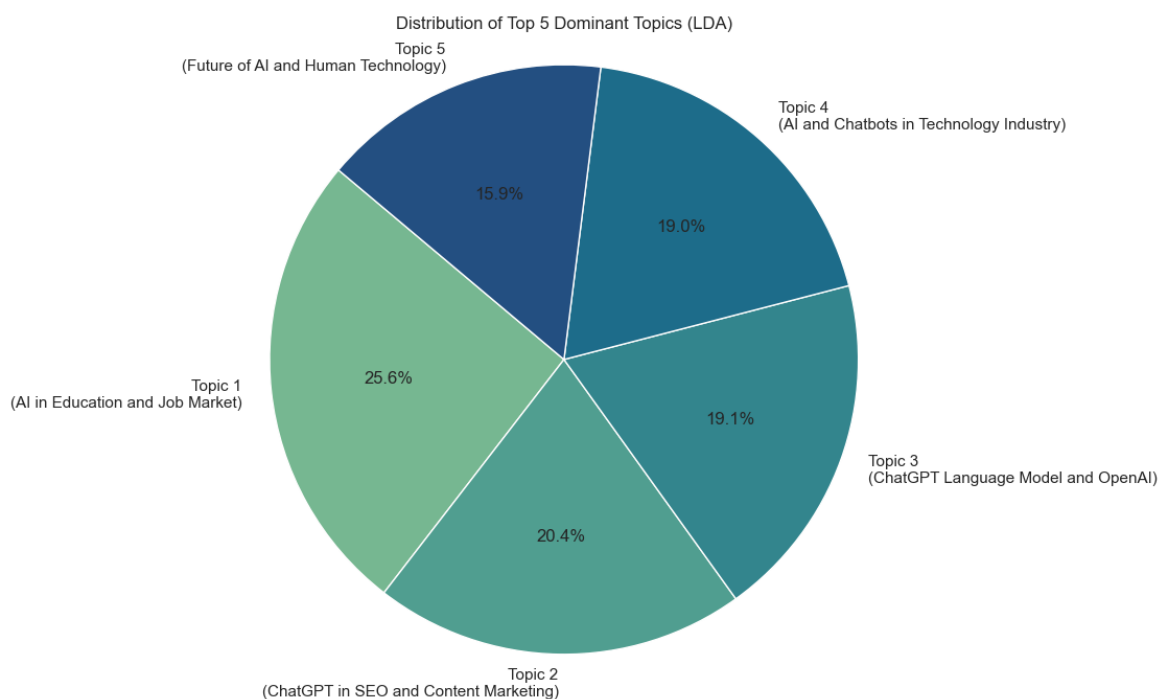


Fig 5: Distribution of Top 5 Dominant Topics

The two topics with the most tweets as observed from Figure 5 and Table 9 Topic 1 ("AI in Education and Job Markets") and Topic 5 ("Future of AI and Human Technology"), together accounted for 40% of all tweets. This demonstrates how widely discussed ChatGPT's effects on employment and education have been.

Another 30% of tweets were on topics with good use cases, such as marketing (Topic 2) and tech talks (Topics 3 and 4). This is consistent with the sentiment analysis-identified, generally positive opinions.

Topic Number	Topic Name	Prevalence
0	AI in Education and Job Market	0.106460
1	ChatGPT in SEO and Content Marketing	0.086707
2	ChatGPT Language Model and OpenAI	0.051701
3	AI and Chatbots in Technology Industry	0.096713
4	Future of AI and Human Technology	0.127061

Table 9: Topic Prevalence for the Top 5 Topics

Topic Number	Topic Name	Positive (%)	Negative (%)	Neutral (%)
0	AI in Education and Job Market	62.66744	17.42912	19.90343
1	ChatGPT in SEO and Content Marketing	64.64902	15.49214	19.85882
2	ChatGPT Language Model and OpenAI	60.96173	12.74710	26.29115
3	AI and Chatbots in Technology Industry	42.11027	16.56119	41.32853
4	Future of AI and Human Technology	56.75696	16.128150	30.42800

Table 10: Sentiment Percentage by Topic for the Topics

The examination of several ChatGPT-related issues gives valuable insights into public mood. Topics 0, 1, and 4 stand out as seen in Table 10:

Topic 0, "AI in Education and Job Market," receives a lot of favorable feedback, demonstrating that people are excited about AI's role in education and job markets. However, a significant proportion expresses unfavorable and neutral opinions, indicating the importance of balanced discourse.

Topic 1, "ChatGPT in SEO and Content Marketing," shares a similar upbeat attitude, emphasizing its potential in digital marketing. Negative emotion is relatively minimal, indicating widespread enthusiasm.

Topic 4, "Future of AI and Human Technology," exhibits a generally favorable tone, emphasizing the optimism surrounding AI's future. However, the unexpectedly large negative sentiment proportion reveals a data anomaly or outlier in this topic.

The results highlight the complexities of public view of AI, with issues 0 and 1 reflecting a positive outlook and topic 4 warranting further examination due to its anomalous traits.

Overall, these results show a positive public response to ChatGPT on Twitter, one that is more concerned with capabilities and possibilities for the future than with hazards. Relevant brands might profit from this enthusiasm and curiosity through marketing. The ramifications of this quickly evolving AI must be addressed by educators in the meantime. It would be wise to conduct further research on potential effects on employment.

4.5 Discussion

This chapter searches into a deep explanation of the findings reported in the preceding chapter. It aims to answer the research questions provided at the commencement of this study, providing insights into public perceptions of ChatGPT, its consequences, and the variables affecting these beliefs.

RQ 1: How can businesses and organizations use the largely positive public reaction about ChatGPT on Twitter to improve their marketing strategies?

In view of ChatGPT's highly positive public opinion, the study of Twitter chats highlights a tremendous opportunity for corporations and organizations. Leveraging this rise in positivity can be critical for marketing initiatives, allowing firms to effectively address consumer queries and improve assistance with ChatGPT-powered chatbots. Furthermore, ChatGPT's content generating capability corresponds with the requirement for new and relevant online content, thereby enhancing SEO performance and user engagement. This positive mood implies a larger willingness to embrace AI technology, allowing enterprises to improve user experiences by incorporating ChatGPT into numerous touchpoints. However, responsible AI use is critical, balancing automation and human interactions while adhering to ethical rules to build customer trust and authenticity.

RQ 2: What implications do public sentiments and perceptions about ChatGPT's role in education and the job market, as revealed in Twitter conversations, have for educational and employment strategies in an AI-driven era?

The analysis emphasizes the critical importance of understanding the potential repercussions of ChatGPT in education and the job market (Topic 0). The public debate on this topic emphasizes the recognition of AI's transformational

potential in these critical sectors. The overall favourable tone of these exchanges suggests a positive outlook on ChatGPT's ability to improve learning experiences and boost job market outcomes. ChatGPT is viewed as a useful resource in education, with the ability to change traditional learning methodologies. Its ability to deliver real-time information and customized help promises to improve student learning and research abilities. Its adaptability to individual learning styles also holds the potential of tailored educational experiences. However, worries have developed amid this promising potential. One major concern is that students would become overly reliant on AI technologies like ChatGPT, which could stifle the development of critical thinking and research abilities. The collection and analysis of large amounts of student data raises privacy and security concerns, needing stringent measures.

Public sentiments remain positive towards Chat GPT. It has the potential to automate routine tasks, thus increasing employee job satisfaction by eliminating repeated work. Nonetheless, the employment market has its own set of difficulties. Task automation by AI, such as ChatGPT, presents the danger of employment displacement in some industries. Workers in professions that rely primarily on routine duties may experience substantial difficulties. Concerns about justice, bias, and discrimination, as well as ethical and regulatory issues, loom big. As a result, there is an increasing demand for legislation to safeguard the ethical use of AI in hiring and HR operations.

ChatGPT has the potential to have a substantial positive impact on education and the employment market by improving learning experiences, simplifying processes, and boosting skill development. However, these advantages must be evaluated against potential disadvantages such as overreliance on technology in education, privacy problems, employment displacement, and ethical concerns.

Future study must continue to investigate these effects thoroughly, with an emphasis on devising techniques to maximize the benefits while limiting the negative effects of ChatGPT in education and the job market.

RQ 3: What ethical considerations and societal problems are related with ChatGPT's expanding significance, and how may these concerns be addressed responsibly?

While public perception is mostly positive, it is critical to acknowledge that AI technology, such as ChatGPT, raises ethical and societal challenges. Users and organizations should be aware of these challenges to ensure responsible AI implementation. Concerns have been raised about data privacy, disinformation dissemination, and employment displacement. To alleviate these issues, AI ethics and governance frameworks should be established and applied. OpenAI and other stakeholders should promote transparency and accountability in AI development and deployment.

To enable users to comprehend the capabilities and limitations of ChatGPT, society must build an AI literacy culture. Awareness campaigns and educational initiatives can assist the public in making educated judgments concerning AI use.

RQ 4: What accounts for the prevalence of both optimism and skepticism in popular perceptions of AI, and what explains the unanticipated rise in negative attitude towards AI's potential societal implications?

The study revealed a rich tapestry of sentiments about AI technologies, including both enthusiasm and skepticism. AI's ability to optimize corporate processes, speed content development, and improve user interactions are among the reasons for optimism. Skepticism, on the other hand, derives from concerns about potential job displacement, ethical quandaries, and the prospect of AI

misuse. Nonetheless, the surprising spike in negative opinion in the "Future of AI and Human Technology" topic warrants further investigation. This oddity necessitates a more in-depth investigation of the underlying causes of intermittent negativity during mainly positive opinions. Investigating this anomaly has the potential to give light on solutions for addressing concerns, encouraging responsible AI research, and encouraging better informed public debates about the future of AI and its societal ramifications.

In conclusion, this study examined into public impressions of ChatGPT on Twitter, revealing a largely positive sentiment about this AI technology. Businesses can use this positivity to better their marketing tactics and consumer experiences. Furthermore, the study offered light on the far-reaching consequences of ChatGPT in education and the labour market, emphasizing both the possible benefits and accompanying difficulties, such as overreliance on technology and ethical quandaries. Ethical concerns and societal difficulties associated to AI's growing importance were addressed, urging the creation of ethical frameworks and enhanced AI literacy. Finally, the unanticipated increase in negative sentiment toward AI's societal ramifications calls for additional research to inform responsible AI development and create informed public conversation.

5. Conclusion

5.1 Research Summary

The goal of this research was to use Twitter data to determine both qualitative and quantitative trends in public opinion and demand for ChatGPT. Sentiment analysis, topic modeling, and exploratory data analysis were used to get important insights into public opinion and perceptions of this contagious AI technology.

A majority of the tweets, or 50% of them, expressed enthusiasm and optimism, according to the sentiment analysis. Less was said critically about hazards and restrictions. Discussions about the effects of AI on employment and education were identified by topic modeling. Use cases for marketing and the future were other important issues. However, there was little discussion of dangers and restrictions. These results imply that ChatGPT is still in a buzz cycle where potential, and capabilities are more important than practicality. A more balanced conversation may develop as firsthand knowledge grows. Tracking this development will require ongoing observation.

The use of topic modeling tools revealed a wide range of discussions and debates. The ongoing debate concerning AI's impact on jobs and education was foremost among these, reflecting broader societal worries. Furthermore, the research revealed debates about ChatGPT's possible marketing applications as well as projections about its future trajectory across many domains.

Examination of the precision and accuracy of sentiment evaluations, thorough evaluation metrics for the different approaches of sentiment analysis, have shown to be invaluable. This methodological rigor makes sure that our sentiment

findings are valid. Additionally, identifying the common themes in the Twitter dialogue has given interpreters of sentiment patterns significant perspective.

As a result, while qualitative signals are encouraging, quantitative analysis highlight the importance of validating social media signals versus real-world traction. For accurate technology forecasting, multiple data sources are required.

5.2 Contributions to the field

This work contributes significantly to the area by undertaking one of the first empirical assessments of public opinion on ChatGPT. It sheds insight on the growing sentiment surrounding this emerging AI technology by leveraging a large dataset of 500,000 tweets, acting as a helpful reference for future research. It also explains how to effectively employ computational approaches to gain strategic insights from social media data connected to future technology. This study helps firms with chatbot business strategy and marketing by giving data-driven insights into customer sentiments and prospective use cases.

However, it is critical to recognize the limitations of social data interpretation due to demographic biases and subjectivity. In summary, this groundbreaking study combines qualitative and quantitative assessments of social media data to shed light on the effects of disruptive AI such as ChatGPT, giving critical insights for technology companies, politicians, and educational institutions.

5.3 Practical Implications

This study shows numerous practical ramifications for technology companies and policymakers.

This study gives useful practical insights for both technology corporations. First and foremost, AI enterprises should be aware of the overwhelming positive

attitude and public interest in AI technology, as these offer enormous growth prospects. Marketing techniques should be developed to safely harness this excitement, but it is critical to strike a balance between promotion and moderating public expectations.

While favourable sentiment indicates eagerness, the report emphasizes the uncertainty around user retention and profitability. Companies should avoid making quick extrapolations based purely on social media trends and instead focus on more complete indicators. It is critical to continuously monitor sentiment and new issues on social media. This enables businesses to quickly spot growing threats, misinformation, and public concerns. Proactive transparency measures can aid in the development of trust and the reduction of potentially negative perceptions.

It is essential to supplement social media insights with user surveys and adoption indicators. This more thorough approach allows organizations to emphasize value and confidence in their AI products and services by providing a deeper understanding of user sentiments and actions.

The impact of AI on knowledge labour needs curriculum changes for instructors. To fully prepare the future workforce, it is critical to include both technical and ethical training.

The impact of AI on knowledge work needs curriculum revisions for educational institutions. It is critical to include both technical and ethical training to appropriately prepare the future generation.

The role of policymakers in defining the AI environment is essential. They must carefully consider the risks and benefits of AI technologies, engaging experts

from all disciplines to design balanced policies that encourage innovation while protecting social interests.

An examination of employment sensitivity to AI displacement is required in the arena of labour policies. To combat potential job disruptions, policymakers should develop methods for skills retraining.

Finally, it is critical to promote collaboration between private companies and governments. This collaboration has the potential to assist the workforce adapt to the evolving AI landscape by creating complementarity between humans and AI systems.

5.4 Limitations

- **Data Sampling Constraints:** It is critical to recognize the limitations of the data acquired via Twitter. In comparison to the average population, Twitter users tend to be younger, urban, and tech-savvy. This demographic bias may result in an incomplete knowledge of popular mood. Future research should strive for larger demographics to improve representativeness.
- **Geographic and linguistic limitations:** The study focused solely on English-language tweets, which may not accurately portray worldwide sentiments. Future study should examine incorporating data from a broader selection of languages and geographic places to improve generalizability.
- **Inadequate Metadata:** While the research concentrated on textual content, including additional metadata such as user age, gender, and geography could provide more thorough insights. This demographic data may enable more representative.
- **Textual Analysis Difficulties:** Analyzing brief, casual tweets can be difficult, especially when attempting to capture nuanced perspectives. NLP

algorithms may have difficulty extracting complex sentiments from such restricted text. Furthermore, the subjectivity involved in evaluating issues modeled using techniques such as LDA presents difficulties. Future research could investigate more advanced NLP models, such as BERT, to better capture semantic links.

- **Scope Restrictions:** The study's timeframe was limited to three months, which may not be enough to correctly examine long-term patterns. Due to volume constraints, some opinions and conversations may have been overlooked. To overcome this, future study should try extending the observation duration and experimenting with different methods of recording specialized discussions.

5.5 Ethical Considerations

This research study demonstrates a strong commitment to moral standards in all its components. First off, it demonstrates a commitment to protecting user privacy that Twitter data is responsibly used rather than private information. Additionally, anonymity and privacy protection are improved by meticulously removing usernames during data preprocessing. Given the presence of demographic biases within the Twitter sample, the interpretation of the results was addressed with caution. This admission emphasizes the need for caution when drawing conclusions from such data. The paper's interpretations of topic modeling showed a dedication to sound research procedures and responsible reporting by considering model tuning restrictions and the need for qualitative human analysis.

The study also argued for the ethical advancement of AI technology, based on an awareness of public opinion. It demanded the creation of moral frameworks

to direct the development of AI while defending the public's interests and fostering inclusive, diverse, and equitable AI systems.

As a result of emphasizing moral issues and ethical research methods, this research study provides an admirable example and advances AI technology in a way that is consistent with society values and broader public interests.

5.6 Future Research

- **Multi-Source Analysis:** To overcome the limits caused by data volume limitations and demographic bias, future research could employ a cross-platform analysis technique. Aside from text-based media like Twitter, modern thoughts and reviews are increasingly being expressed in video formats on sites like YouTube, Instagram, and TikTok. While expanding into these multimedia areas may need additional resources, it promises a more comprehensive and nuanced picture of public sentiment. Researchers can acquire a more holistic view of the ever-changing landscape of public thoughts and sentiments by including diverse platforms such as Reddit, forums, blogs, and multimedia-sharing networks.
- **In-Depth Qualitative Research:** Surveys, interviews, and focus groups can supplement quantitative social media analysis by offering more in-depth viewpoints and context. Qualitative research approaches can assist in eliciting subtle feelings and motivations.
- **Long-Term Trend Analysis:** Future research should assess patterns over years to understand how sentiments shift over time, considering the dynamic nature of AI technologies and the competitive landscape.
- **Human-AI Collaboration Research:** Examining how humans engage and work with AI systems in a variety of sectors, such as healthcare, education,

and customer service, might provide insights into the practical implications of AI adoption.

- **Initiatives in Education:** Future work could include the creation of AI education programs that address both technological and ethical concerns. These programs can be tailored to various age groups and educational levels.
- **Public Policy study:** Researchers might concentrate on the development and study of AI-related public policy. This could include assessing the efficacy of existing restrictions or suggesting new policies that strike a balance between innovation and public well-being.
- **Collaboration with Industry:** By collaborating with AI firms and organizations for data sharing and cooperative research initiatives, more thorough and meaningful studies can be conducted. This will necessitate forming alliances with key stakeholders.

These future study directions can build on the dissertation report's basis, allowing for a more in-depth and comprehensive understanding of the complicated interaction between AI technology and public sentiment, as well as their broader societal ramifications. This will provide more robust, validated, and actionable information into how AI technologies like ChatGPT are affecting the public. The integrated approach highlights the relevance of social media analytics while emphasizing the importance of multidimensional assessment. As AI spreads over the world, ongoing interdisciplinary research is vital.

References

1. 500k ChatGPT-related Tweets Jan-Mar 2023. (n.d.). Wwww.kaggle.com.
<https://www.kaggle.com/datasets/khalidryder777/500k-chatgpt-tweets-jan-mar-2023>
2. Global chatbot market value 2018-2027. (n.d.). Statista.
<https://www.statista.com/statistics/1007392/worldwide-chatbot-market-size/>
3. Feng, H. (2023, August 2). How AI-powered chatbots are transforming marketing and sales operations. IBM Blog.
<https://www.ibm.com/blog/how-ai-powered-chatbots-are-transforming-marketing-and-sales-operations/>
4. Carpenter, T. A. (2023, January 11). @TAC_NISO questions ChatGPT on scholarly communications. The Scholarly Kitchen.
<https://scholarlykitchen.sspnet.org/2023/01/11/chatgpt-thoughts-on-ais-impact-on-scholarly-communications/?informz=>
5. Araujo, T. (2018). Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. Computers in Human Behavior, 85, 183–189. <https://doi.org/10.1016/j.chb.2018.03.051>
6. Business Insider Intelligence. (2016, December 14). 80% of businesses want chatbots by 2020 - Business Insider. Business Insider; Business Insider.
<https://www.businessinsider.com/80-of-businesses-want-chatbots-by-2020-2016-12?IR=T>
7. Dantas, A. C. (2023, January 5). Ethics in chatGPT and other AI's. Medium.
<https://medium.com/@adilmarcoelhodantas/ethics-in-chatgpt-and-other-ais-ee31ce8e9f09>

8. AlAfnan, M. A., Samira Dishari, Marina Jovic, & Koba Lomidze. (2023). ChatGPT as an Educational Tool: Opportunities, Challenges, and Recommendations for Communication, Business Writing, and Composition Courses. *Journal of Artificial Intelligence and Technology*, 3(2). <https://doi.org/10.37965/jait.2023.0184>
9. Yang, M. (2023, January 6). New York City schools ban AI chatbot that writes essays and answers prompts. *The Guardian*. <https://www.theguardian.com/us-news/2023/jan/06/new-york-city-schools-ban-ai-chatbot-chatgpt>
10. Dare to evolve: Reassessing assessments in business schools. (2022, December 20). Chartered Association of Business Schools. <https://charteredabs.org/dare-to-evolve-re-assessing-assessments-in-business-schools/>
11. Ortiz, S. (2023, April 18). What is ChatGPT and why does it matter? Here's everything you need to know. *ZDNET*. <https://www.zdnet.com/article/what-is-chatgpt-and-why-does-it-matter-heres-everything-you-need-to-know/>
12. ChatGPT News, Research and Analysis. (2023, August 23). *The Conversation*. <https://theconversation.com/uk/topics/chatgpt-130961>
13. mollick, E. (2022, December 14). ChatGPT Is a Tipping Point for AI. *Harvard Business Review*. <https://hbr.org/2022/12/chatgpt-is-a-tipping-point-for-ai>
14. Hulick, K. (2023, April 12). How ChatGPT and similar AI will disrupt education. *Science News*. <https://www.sciencenews.org/article/chatgpt-ai-artificial-intelligence-education-cheating-accuracy>
15. Roose, K. (2023, February 3). How ChatGPT Kicked Off an A.I. Arms Race. *The New York Times*. <https://www.nytimes.com/2023/02/03/technology/chatgpt-openai-artificial-intelligence.html>
16. The Learning Network. (2023, February 2). What Students Are Saying About ChatGPT. *The New York Times*. <https://www.nytimes.com/2023/02/02/learning/students-chatgpt.html>

17. Nast, C. (2023, April 13). *What Kind of Mind Does ChatGPT Have?* The New Yorker. <https://www.newyorker.com/science/annals-of-artificial-intelligence/what-kind-of-mind-does-chatgpt-have>
18. Parkhill, B. (2023, February 1). *What Is ChatGPT? How AI Is Transforming Multiple Industries.* Forbes. <https://www.forbes.com/sites/qai/2023/02/01/what-is-chatgpt-how-ai-is-transforming-multiple-industries/>
19. Wang, F.-Y., Miao, Q., Li, X., Wang, X., & Lin, Y. (2023). What Does ChatGPT Say: The DAO from Algorithmic Intelligence to Linguistic Intelligence. *IEEE/CAA Journal of Automatica Sinica*, 10(3), 575–579. <https://doi.org/10.1109/JAS.2023.123486>
20. Emenike, M. E., & Emenike, B. U. (2023). Was This Title Generated by ChatGPT? Considerations for Artificial Intelligence Text-Generation Software Programs for Chemists and Chemistry Educators. *Journal of Chemical Education*. <https://doi.org/10.1021/acs.jchemed.3c00063>
21. Kooli, C. (2023). Chatbots in Education and Research: A Critical Examination of Ethical Implications and Solutions. *Sustainability*, 15(7), 5614. <https://doi.org/10.3390/su15075614>
22. Biswas, S. S. (2023). Role of Chat GPT in Public Health. *Annals of Biomedical Engineering*, 51. <https://doi.org/10.1007/s10439-023-03172-7>
23. Tlili, A., Shehata, B., Adarkwah, M. A., Bozkurt, A., Hickey, D. T., Huang, R., & Agyemang, B. (2023). What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learning Environments*, 10(1). <https://doi.org/10.1186/s40561-023-00237-x>
24. Wang, F.-Y., Li, J., Qin, R., Zhu, J., Mo, H., & Hu, B. (2023). ChatGPT for Computational Social Systems: From Conversational Applications to Human-Oriented Operating Systems. *IEEE Transactions on Computational Social Systems*, 10(2), 414–425. <https://doi.org/10.1109/tcss.2023.3252679>

25. ChatGPT in Education: Global Reactions to AI Innovations. (2023, May 10).
Www.researchsquare.com. <https://www.researchsquare.com/article/rs-2840105/v1>
26. Javaid, M., Haleem, A., Ravi Pratap Singh, Khan, S., & Ibrahim Haleem Khan. (2023). Unlocking the opportunities through ChatGPT Tool towards ameliorating the education system. 100115–100115.
<https://doi.org/10.1016/j.tbench.2023.100115>
27. Chen, B., Wu, Z., & Zhao, R. (2023). From fiction to fact: the growing role of generative AI in business and finance. Journal of Chinese Economic and Business Studies, 1–26. <https://doi.org/10.1080/14765284.2023.2245279>
28. Holt-Nguyen, C. (2023, May 25). Lessons Learned: The Rise of ChatGPT For Sentiment Analysis. Medium. <https://levelup.gitconnected.com/lessons-learned-the-rise-of-chatgpt-for-sentiment-analysis-13e7debd6cbd>
29. Paiva, F. C. L. (2023, April 26). Can ChatGPT Compete with Domain-Specific Sentiment Analysis Machine Learning Models? Medium.
<https://towardsdatascience.com/can-chatgpt-compete-with-domain-specific-sentiment-analysis-machine-learning-models-cdcd9937b460>
30. Araujo, A. F., Gôlo, M. P. S., & Marcacini, R. M. (2021). Opinion mining for app reviews: an analysis of textual representation and predictive models. Automated Software Engineering, 29(1). <https://doi.org/10.1007/s10515-021-00301-1>
31. Rouhani, S., & Mozaffari, F. (2022). Sentiment analysis researches story narrated by topic modeling approach. Social Sciences & Humanities Open, 6(1), 100309. <https://doi.org/10.1016/j.ssaho.2022.100309>
32. Holt-Nguyen, C. (2023, May 16). Compare ChatGPT to Machine Learning Techniques for Sentiment Analysis in 2023. Medium.
<https://pub.towardsai.net/compare-chatgpt-to-machine-learning-techniques-for-sentiment-analysis-in-2023-3e897fc22da1>

33. Venkatakrishnan, S., Kaushik, A., & Verma, J. K. (2020). Sentiment Analysis on Google Play Store Data Using Deep Learning. *Algorithms for Intelligent Systems*, 15–30. https://doi.org/10.1007/978-981-15-3357-0_2
34. Mujahid, M., Kanwal, K., Furqan Rustam, Wajdi Aljadani, & Ashraf, I. (2023). Arabic ChatGPT Tweets Classification using RoBERTa and BERT Ensemble Model. <https://doi.org/10.1145/3605889>
35. George, A. S., George, A. H., & Martin, A. G. (2023). A Review of ChatGPT AI's Impact on Several Business Sectors. *Partners Universal International Innovation Journal (PUIJ)*, 01 (01), 9–23.
36. Reddy, P. (n.d.). Understanding the Power of ChatGPT's Sentiment Analysis for Customer Feedback Analysis. *Www.c-Sharpcorner.com*. Retrieved September 4, 2023, from <https://www.c-sharpcorner.com/article/understanding-the-power-of-chatgpts-sentiment-analysis-for-customer-feedback-an/>
37. Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations, and future scope. *Internet of Things and Cyber-Physical Systems*, 3, 121–154. <https://doi.org/10.1016/j.iotcps.2023.04.003>
38. Frąckiewicz, M. (2023, April 18). The Role of ChatGPT in Enhancing Sentiment Analysis for News and Media Monitoring. *TS2 SPACE*. <https://ts2.space/en/the-role-of-chatgpt-in-enhancing-sentiment-analysis-for-news-and-media-monitoring/>
39. Alamanda, D. T., Ramdhani, A., Kania, I., Susilawati, W., & Hadi, E. S. (2019). Sentiment analysis using text mining of Indonesia tourism reviews via social media. *Int. J. Humanit. Arts Soc. Sci*, 5(2), 72-82.
40. Thakur, O., Sri Khetwat Saritha, & Jain, S. (2022). Topic Modeling, Sentiment Analysis and Text Summarization for Analyzing News Headlines and Articles. *Communications in Computer and Information Science*, 220–239. https://doi.org/10.1007/978-3-031-24352-3_18

41. Li, S., Xie, Z., Dickson K.W. Chiu, & Ho, K. (2023). Sentiment Analysis and Topic Modeling Regarding Online Classes on the Reddit Platform: Educators versus Learners. *Applied Sciences*, 13(4), 2250–2250. <https://doi.org/10.3390/app13042250>
42. How good is ChatGPT 4? We analyse its answers to four questions... (2023, June 27). GWS Media. <https://www.gwsmedia.com/articles/how-good-is-chat-gpt-4>
43. Josh Bersin. (2023, January 22). Understanding Chat-GPT, And Why It's Even Bigger Than You Think. JOSH BERSIN. <https://joshbersin.com/2023/01/understanding-chat-gpt-and-why-its-even-bigger-than-you-think/>
44. Krishnamurthy, S. (2023, February 22). An Analysis of ChatGPT and OpenAI GPT-3: How to Use it For Your Business. Version 1. <https://www.version1.com/an-analysis-of-chatgpt-and-openai-gpt3-how-to-use-it-for-your-business/>
45. Ruby, M. (2023, January 31). How ChatGPT Works: The Models Behind The Bot. Medium. <https://towardsdatascience.com/how-chatgpt-works-the-models-behind-the-bot-1ce5fca96286>
46. Hetler, A. (2023, March). What is ChatGPT? Everything You Need to Know. WhatIs.com. <https://www.techtarget.com/whatis/definition/ChatGPT>
47. Yu, H. (2023). Reflection on whether Chat GPT should be banned by academia from the perspective of education and teaching. *Frontiers in Psychology*, 14. <https://doi.org/10.3389/fpsyg.2023.1181712>
48. Pop!, P. (2023, March 25). History and Future Impact of Chat GPT. Pop! Automation. <https://www.popautomation.com/post/history-and-impact-of-chat-gpt>
49. Chat GPT Negative Impact: Destroying Students and Their Capacity for Development - Educational Documentary & Media Site - ABIS MEDIA. (2023, May 14). Abismedia.com. <https://abismedia.com/chat-gpt-negative-impact/>
50. The Impact of Chat GPT on Education: The Good and the Bad. (2023, July 4). Digital Learning Institute. <https://www.digitallearninginstitute.com/blog/the-impact-of-chat-gpt-on-education/>
51. Varwandkar, A. (2023, February 10). Is Chat GPT the end of thinking skills? The Times of India. <https://timesofindia.indiatimes.com/blogs/the-next-step/is-chat-gpt-the-end-of-thinking-skills/>

52. Editorial Desk. (2023, April 2). Tech Business News. Tech Business News. <https://www.techbusinessnews.com.au/blog/chatgpt-may-lead-to-the-downfall-of-eduction-and-critical-thinking/>
53. Guadalupe, A., Jackeline, G., Pedro, J., Quispe, H., Antonio, M., Yanowsky, G., Ricardo, H., Marina, R., Victor, H., & Arias-González, L. (2023). Effect of Chat GPT on the digitized learning process of university students. 33, 1–15. <https://doi.org/10.59670/jns.v33i.411>
54. IBM100 - Deep Blue. (2012, March 7). Wwww-03.ibm.com. <https://www.ibm.com/ibm/history/ibm100/us/en/icons/deepblue/impact/s/>
55. Saini, A. (2021, July 3). Conceptual Understanding of Logistic Regression for Data Science Beginners [Review of Conceptual Understanding of Logistic Regression for Data Science Beginners]. <https://www.analyticsvidhya.com/blog/2021/08/conceptual-understanding-of-logistic-regression-for-data-science-beginners/>
56. Ray, S. (2017, September 11). Naive Bayes Classifier Explained: Applications and Practice Problems of Naive Bayes Classifier [Review of Naive Bayes Classifier Explained: Applications and Practice Problems of Naive Bayes Classifier]. <https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/>
57. Grand View Research. (2014). Chatbot Market Size And Share Analysis | Industry Report, 2014 - 2025. Grandviewresearch.com. <https://www.grandviewresearch.com/industry-analysis/chatbot-market>
58. Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval*, 2(1–2), 1-135.
59. Hutto, C., & Gilbert, E. (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Proceedings of the international AAAI conference on web and social media (Vol. 8, No. 1, pp. 216-225).
60. Manyika, J., Lund, S., Chui, M., Bughin, J., Woetzel, J., Batra, P., Ko, R., & Sanghvi, S. (2017). Jobs lost, jobs gained: What the future of work will mean for jobs, skills, and wages. McKinsey & Company. <https://www.mckinsey.com/featured-insights/future-of-work/jobs-lost-jobs-gained-what-the-future-of-work-will-mean-for-jobs-skills-and-wages>
61. Adepoju, O. (2021). Reskilling for Construction 4.0. Re-Skilling Human Resources for Construction 4.0, 197–219. https://doi.org/10.1007/978-3-030-85973-2_9
62. Davenport, T. H., & Kirby, J. (2016). Only humans need apply: Winners and losers in the age of smart machines. New York: Harper Business.

63. Asante, I. O., Jiang, Y., Hossin, A. M., & Luo, X. (2023). OPTIMIZATION OF CONSUMER ENGAGEMENT WITH ARTIFICIAL INTELLIGENCE ELEMENTS ON ELECTRONIC COMMERCE PLATFORMS. *Journal of Electronic Commerce Research*, 24(1), 7-28.
64. Saini, M., Arora, V., Singh, M., Singh, J., & Adebayo, S. O. (2022). Artificial intelligence inspired multilanguage framework for note-taking and qualitative content-based analysis of lectures. *Education and Information Technologies*. <https://doi.org/10.1007/s10639-022-11229-8>
65. Wang, X., Zhu, W., & Wang, W. Y. (2023). Large language models are implicitly topic models: Explaining and finding good demonstrations for in-context learning. *arXiv preprint arXiv:2301.11916*.
66. Zhang, Y., Lin, H., Wang, Y., & Fan, X. (2023). Sinophobia was popular in Chinese language communities on Twitter during the early COVID-19 pandemic. *Humanities and Social Sciences Communications*, 10(1), 1–12. <https://doi.org/10.1057/s41599-023-01959-6>
67. Lai, P. (2017). The Literature Review of Technology Adoption Models and Theories for the Novelty Technology. *Journal of Information Systems and Technology Management*, 14(1), 21–38. <https://doi.org/10.4301/s1807-17752017000100002>
68. Gohar, U., Biswas, S., & Rajan, H. (2023, May 1). Towards Understanding Fairness and its Composition in Ensemble Machine Learning. *IEEE Xplore*. <https://doi.org/10.1109/ICSE48619.2023.00133>
69. Kousiouris, G., Akbar, A., Sancho, J., Ta-shma, P., Psychas, A., Kyriazis, D., & Varvarigou, T. (2018). An integrated information lifecycle management framework for exploiting social network data to identify dynamic large crowd concentration events in smart cities applications. *Future Generation Computer Systems*, 78, 516–530. <https://doi.org/10.1016/j.future.2017.07.026>
70. Debnath, S., Seth, D., Sourish Pramanik, Adhikari, S., Mondal, P., Sherpa, D., Sen, D., Mukherjee, D., & Mukerjee, N. (2022). A comprehensive review and meta-analysis of recent advances in biotechnology for plant virus research and significant accomplishments in human health and the pharmaceutical industry. *Biotechnology & Genetic Engineering Reviews*, 1–33. <https://doi.org/10.1080/02648725.2022.2116309>
71. Hartmann, J., & Netzer, O. (2023). Natural language processing in marketing. In *Artificial Intelligence in Marketing* (Vol. 20, pp. 191-215). Emerald Publishing Limited.
72. Huang, S. H., Tsao, S. F., Chen, H., Bin Noon, G., Li, L., Yang, Y., & Butt, Z. A. (2022). Topic modelling and sentiment analysis of tweets related to

- freedom convoy 2022 in Canada. *International journal of public health*, 67, 1605241.
73. Arif Jetha, Hamid Reza Bakhtari, Rosella, L., Monique, Biswas, A., Faraz Vahid Shahidi, Smith, B. T., Smith, M. J., Mustard, C., Khan, N., Arrandale, V. H., Peter John Loewen, Zuberi, D., Dennerlein, J. T., Bonaccio, S., Wu, N., Irvin, E., & Smith, P. (2023). Artificial intelligence and the work–health interface: A research agenda for a technologically transforming world of work. *American Journal of Industrial Medicine*.
<https://doi.org/10.1002/ajim.23517>
 74. Li, S., & Chen, Y. (2022). How non-fungible tokens empower business model innovation. *Business Horizons*.
<https://doi.org/10.1016/j.bushor.2022.10.006>
 75. Kooli, C. (2023). Chatbots in Education and Research: A Critical Examination of Ethical Implications and Solutions. *Sustainability*, 15(7), 5614. <https://doi.org/10.3390/su15075614>
 76. Cecilia. (2023). A comprehensive AI policy education framework for university teaching and learning. *International Journal of Educational Technology in Higher Education*, 20(1). <https://doi.org/10.1186/s41239-023-00408-3>
 77. Lanz, L., Briker, R., & Gerpott, F. H. (2023). Employees Adhere More to Unethical Instructions from Human Than AI Supervisors: Complementing Experimental Evidence with Machine Learning.
<https://doi.org/10.1007/s10551-023-05393-1>
 78. Menczer, F., Crandall, D., Ahn, Y.-Y., & Kapadia, A. (2023). Addressing the harms of AI-generated inauthentic content. *Nature Machine Intelligence*, 5(7), 679–680. <https://doi.org/10.1038/s42256-023-00690-w>
 79. R, S., Mujahid, M., Rustam, F., Shafique, R., Chunduri, V., Villar, M. G., Ballester, J. B., Diez, I. de la T., & Ashraf, I. (2023). Analyzing Sentiments Regarding ChatGPT Using Novel BERT: A Machine Learning Approach. *Information*, 14(9), 474. <https://doi.org/10.3390/info14090474>
 80. Saheb, T., Saheb, T., & Carpenter, D. O. (2021). Mapping research strands of ethics of artificial intelligence in healthcare: A bibliometric and content analysis. *Computers in Biology and Medicine*, 135, 104660.
<https://doi.org/10.1016/j.combiomed.2021.104660>
 81. Saini, R., Mussbacher, G., Guo, J. L. C., & Kienzle, J. (2022). Automated, interactive, and traceable domain modelling empowered by artificial intelligence. *Software and Systems Modeling*, 21(3), 1015–1045.
<https://doi.org/10.1007/s10270-021-00942-6>

Appendix

Appendix A: List of Abbreviations

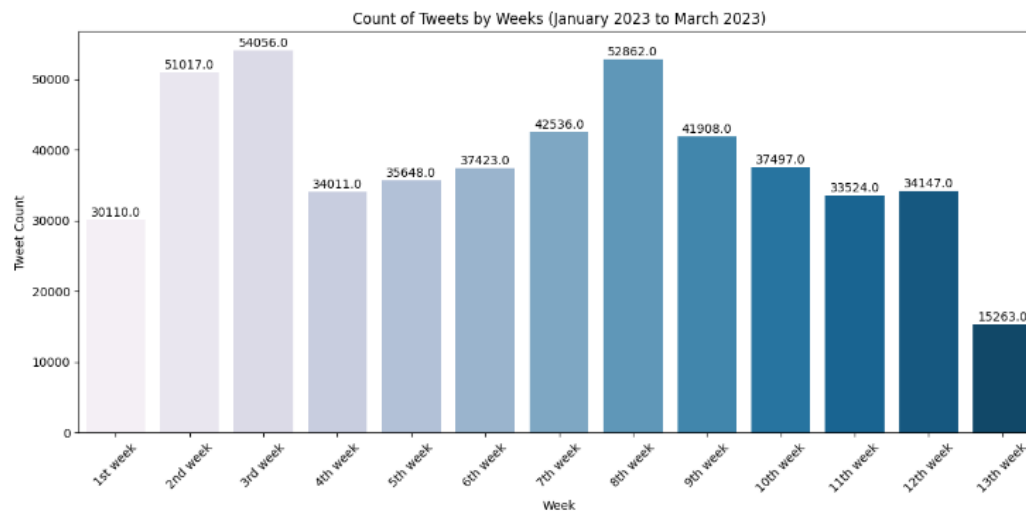
Appendix B: Additional Figures from Analyses

Appendix C: Programming Python Code (Jupyter Labs)

Appendix A: List of Abbreviations

Abbreviation	Full Form
RQ	Research Question
CAGR	Compound Annual Growth Rate
WEF	World Economic Forum
NLP	Natural Language Processing
LDA	Latent Dirichlet Allocation
AI	Artificial Intelligence

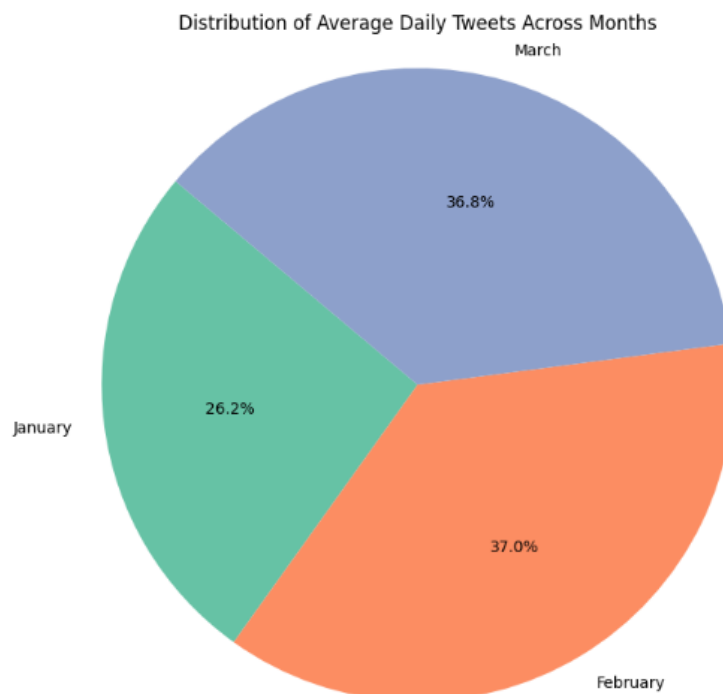
Appendix B: Additional Figures from Analyses



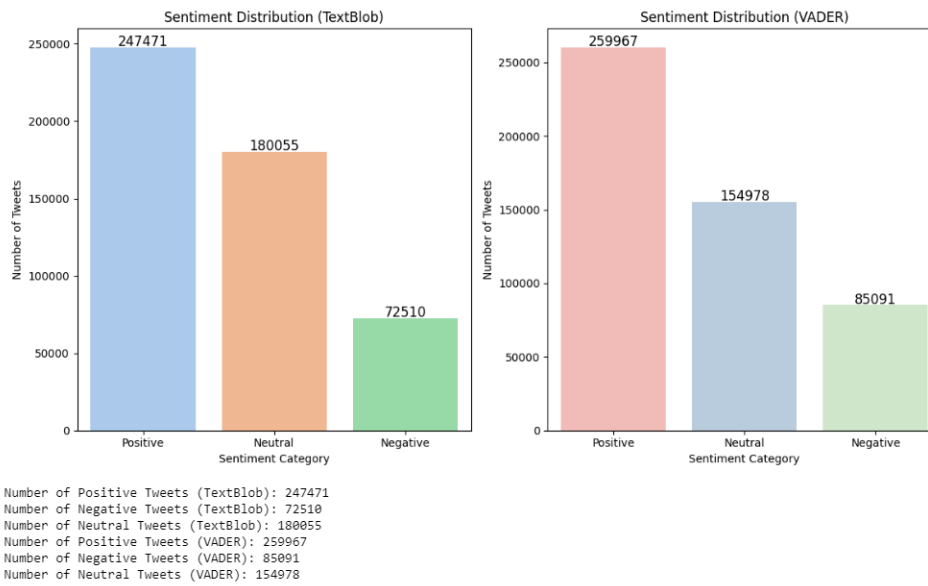
The tweet activity from the 1st week to the 4th week increased by 0.41 times.
The tweet activity from the 4th week to the 8th week increased by 1.0 times.
The tweet activity from the 8th week to the last week increased by 1.24 times.

The above figure depicts Count of Tweets by Weeks (January 2023 to March 2023).

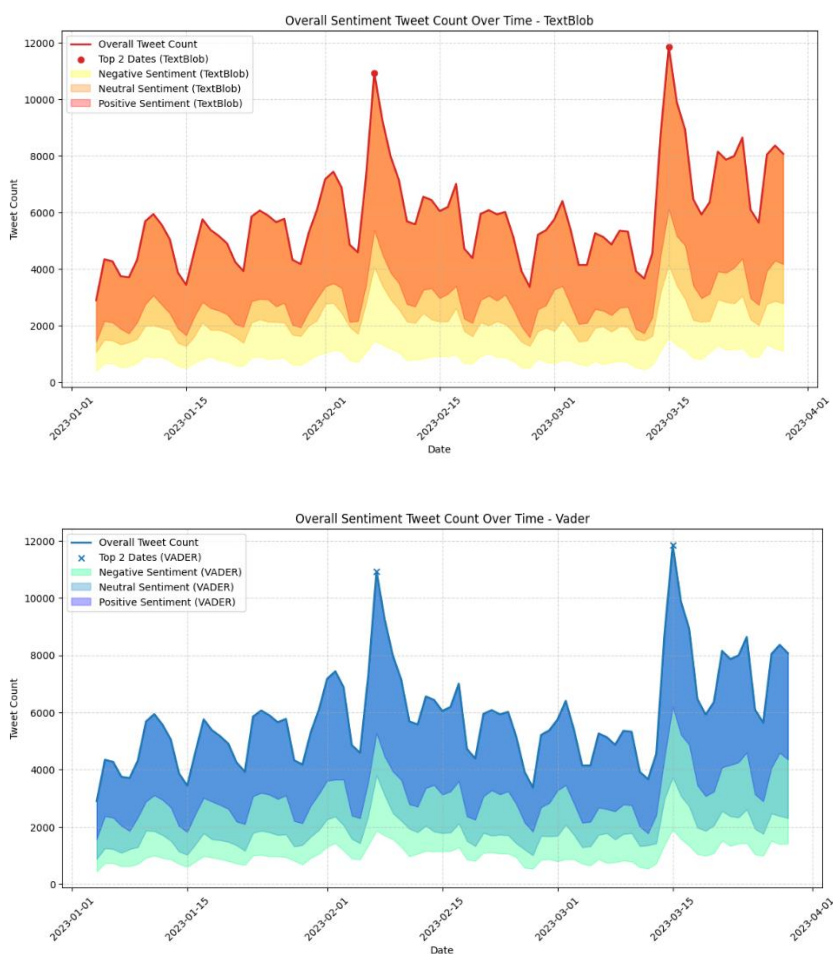
Average daily tweets in January, 2023: 4386.87
Average daily tweets in February, 2023: 6183.00
Average daily tweets in March, 2023: 6157.58



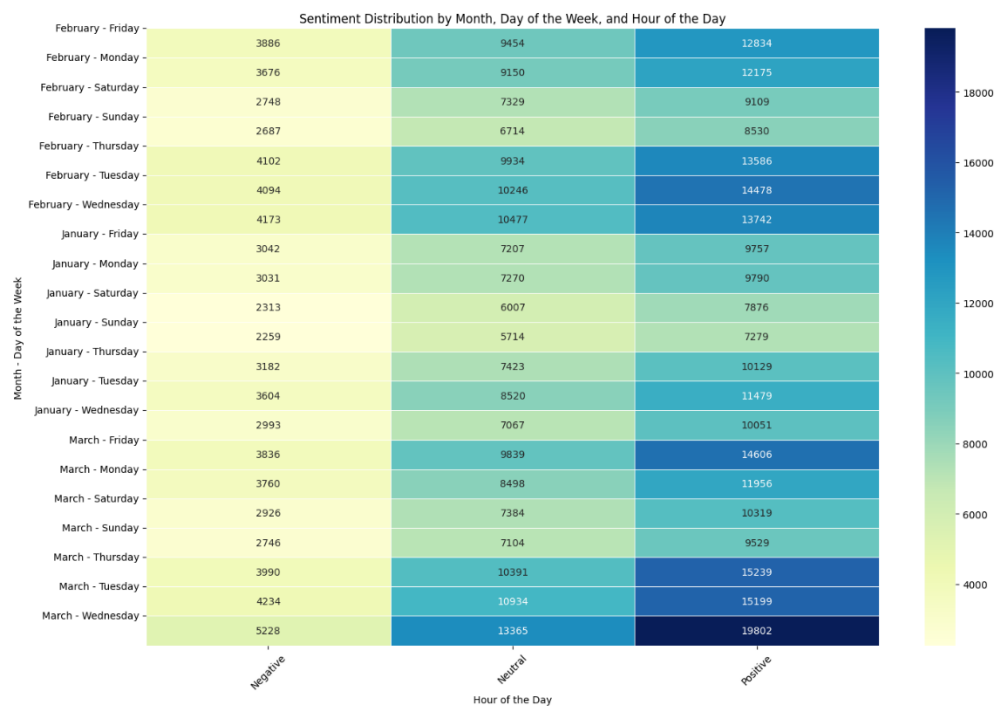
This figure shows the Distribution of Average Daily Tweets Across Months.



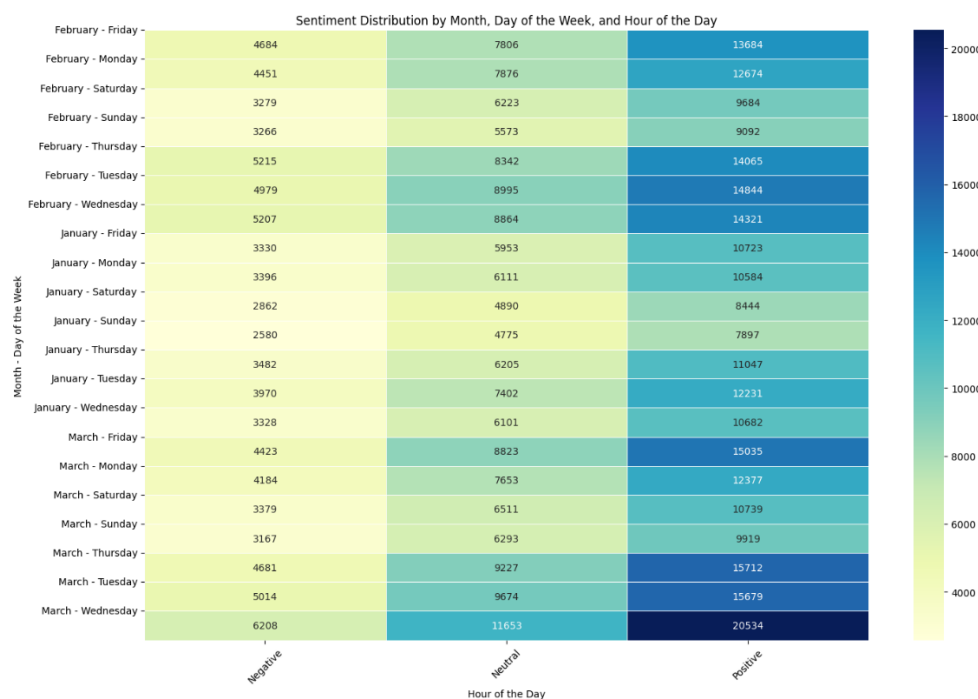
The figure above shows the number of tweets for each type of the sentiment type using both Textblob and VADER.



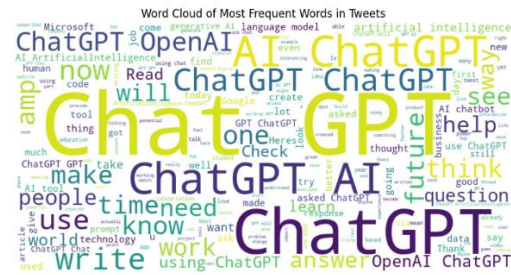
The above figures show the overall sentiment tweet calculated at every 15 day interval using VADER and Textblob.



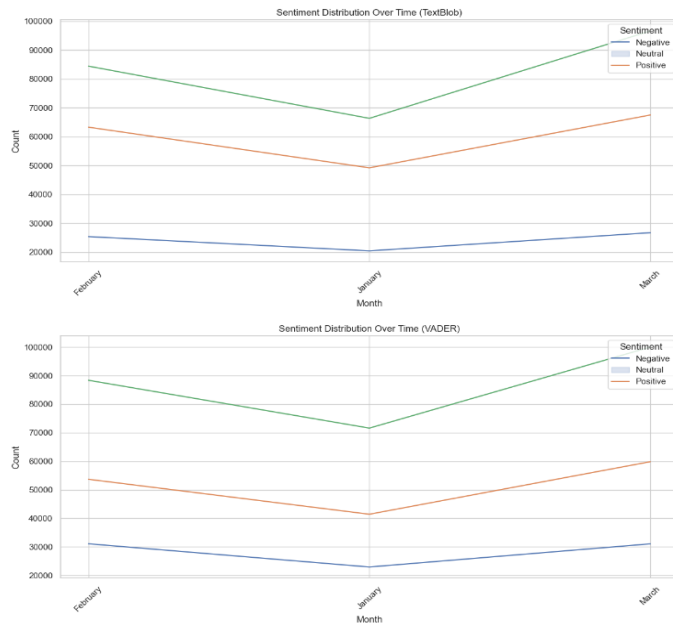
This figure shows sentiment heatmap using Textblob.



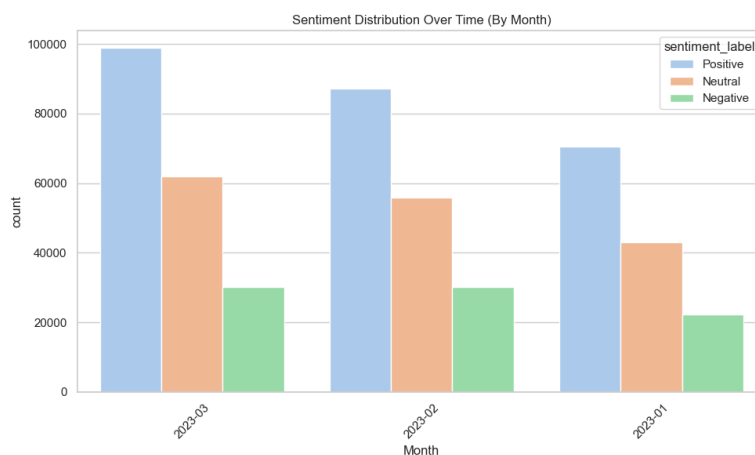
This figure shows sentiment heatmap using VADER.



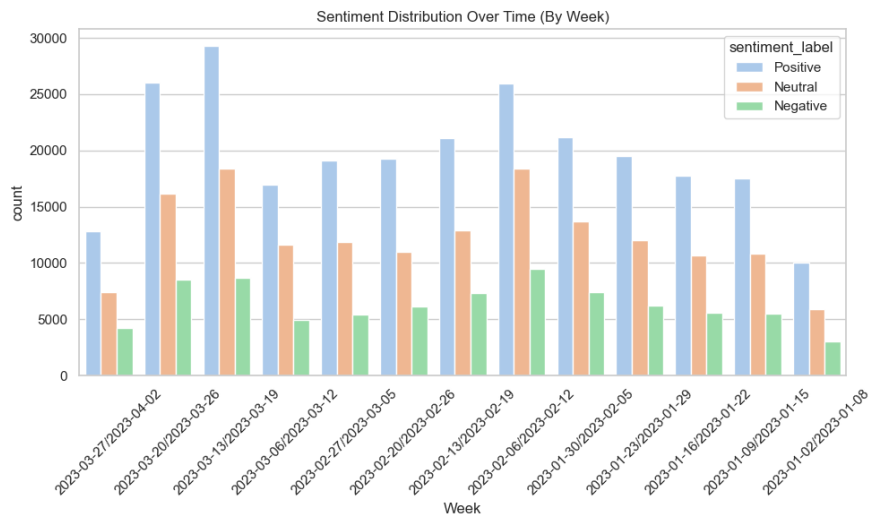
The above figure depicts the most frequent words found in the dataset.



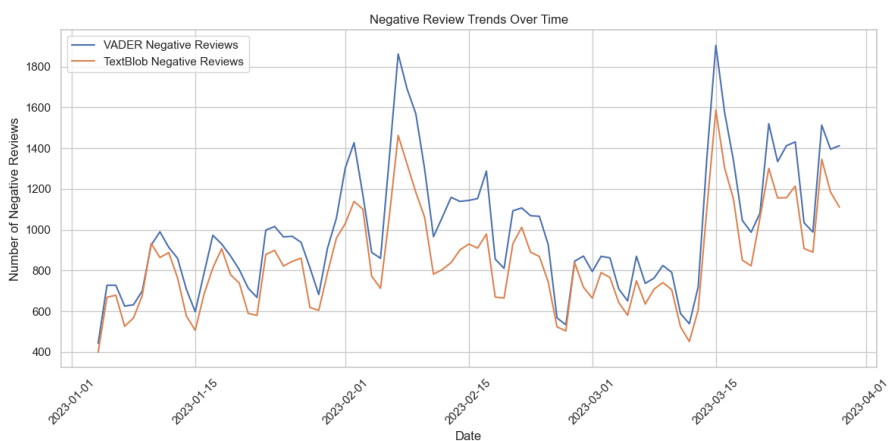
The above figure shows the sentiment Distribution over time using both Textblob and VADER.



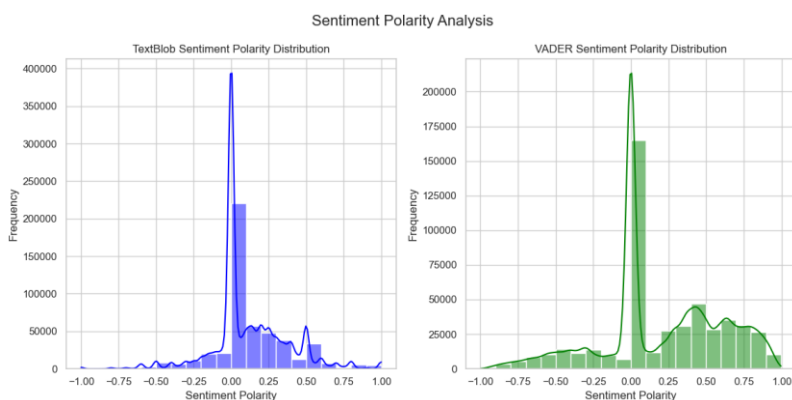
This image shows Sentiment distribution over time by month using the general sentiment label.



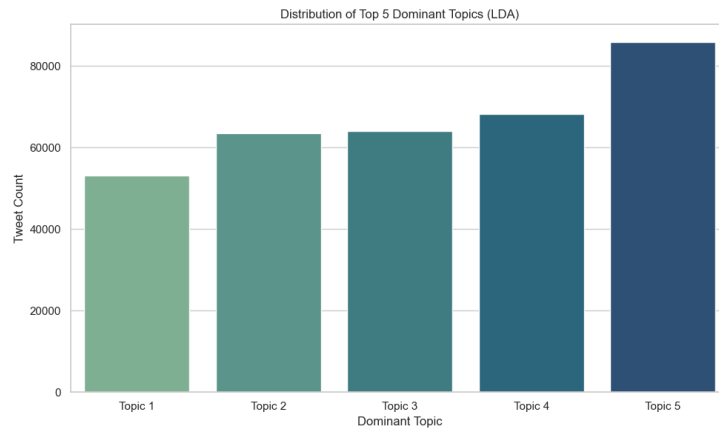
The above figure shows Sentiment distribution over time by each week using the general sentiment label.



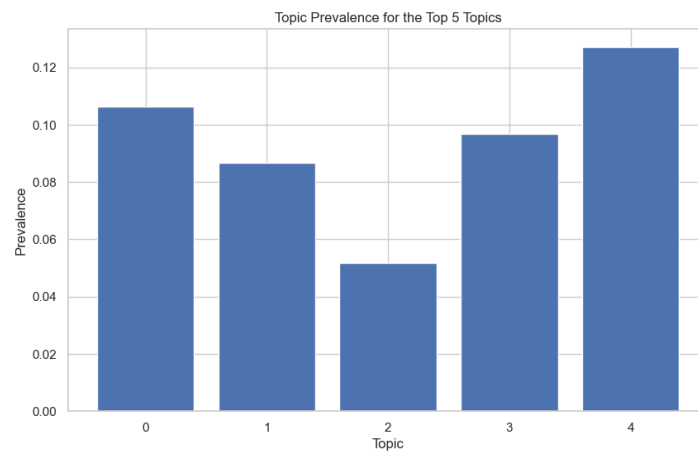
The above figure depicts Negative Review Trends Over Time for both VADER and Textblob.



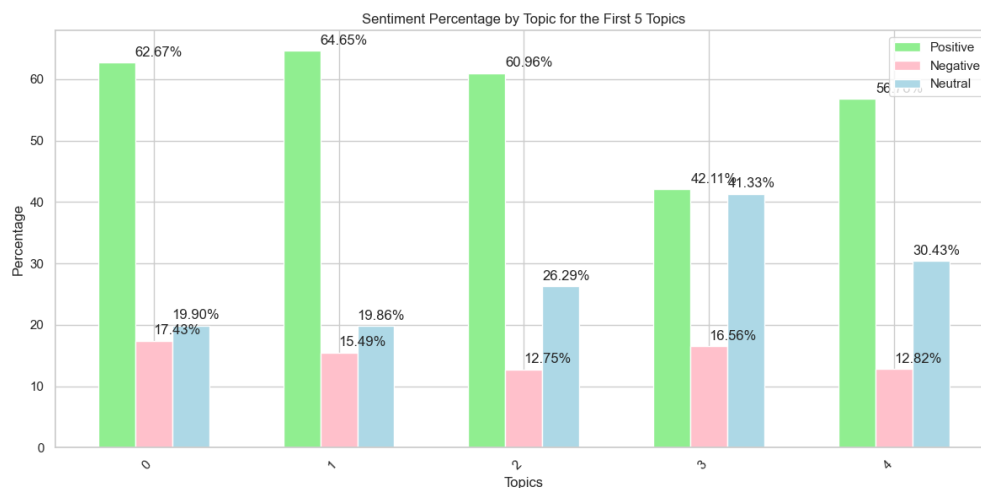
The above figure shows the calculated sentiment polarity for both the methods.



The above figure shows the Distribution of Top 5 Dominant Topics using LDA.



The above figure shows the Topic Prevalence for the Top 5 Topics.



The above figure shows Sentiment Percentage by Topic for the First 5 Topics for positive, negative and neutral sentiments.

Appendix C: Programming Python Code

The code was written in jupyter notebook.

One drive link is provided below if required.

https://universityofexeteruk-my.sharepoint.com/:u:/g/personal/an576_exeter_ac_uk/EYGhT8AgcbVLic5_FGMbRtgBiRsuQ63K8lGb6dYYW4oBmA?e=Dra8GV

```
# Importing Required Libraries
```

```
import pandas as pd
```

```
import numpy as np
```

```
import calendar
```

```
import re
```

```
import nltk
```

```
from nltk.corpus import stopwords
```

```
from nltk.tokenize import word_tokenize
```

```
from nltk.stem import PorterStemmer, WordNetLemmatizer
```

```
import matplotlib.pyplot as plt
```

```
import seaborn as sns
```

```
import plotly.graph_objs as go
```

```
import string
```

```
# Download NLTK resources
```

```
nltk.download('punkt')
```

```
nltk.download('stopwords')
```

```
nltk.download('wordnet')
```

```
from textblob import TextBlob
```

```
from nltk.sentiment import SentimentIntensityAnalyzer
```

```
from datetime import timedelta
```

```
from sklearn.feature_extraction.text import CountVectorizer
```

```
from sklearn.decomposition import LatentDirichletAllocation
```

```
from collections import Counter
```

```
from sklearn.model_selection import train_test_split
```

```

from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score

from sklearn.naive_bayes import MultinomialNB

from sklearn.metrics import classification_report, accuracy_score

from sklearn.metrics import confusion_matrix

from math import sqrt

import matplotlib.dates as mpl_dates

import warnings

from statsmodels.tools.sm_exceptions import ValueWarning

import gensim

from gensim.models import Phrases

from gensim.models.phrases import Phraser

from sklearn.linear_model import LogisticRegression

from sklearn.naive_bayes import MultinomialNB

from sklearn.metrics import roc_curve, roc_auc_score

from sklearn.metrics import mean_squared_error


df_twitter = pd.read_csv('Twitter Jan Mar.csv')


column_names = df_twitter.columns

print("Column names:", column_names)


# Check for empty or blank values in the DataFrame

empty_values = df_twitter[df_twitter['content'].isnull() | (df_twitter['content'] == "")]


# Count the number of empty values

num_empty_values = len(empty_values)


# Print the number of empty values

print("Number of empty values:", num_empty_values)


# Define the placeholder to replace missing content

```

```

placeholder = "No content available"

# Check for null values
if df_twitter.isnull().any().any():
    print("DataFrame contains null values.")

# Replace missing data
df_twitter['content'].fillna(placeholder, inplace=True) # Replace missing content with the
placeholder

# Check for empty or blank values in the DataFrame
empty_values = df_twitter[df_twitter['content'].isnull() | (df_twitter['content'] == "")]

# Count the number of empty values
num_empty_values = len(empty_values)

# Print the number of empty values
print("Number of empty values:", num_empty_values)

#separating time and date
df_twitter['date'] = pd.to_datetime(df_twitter['date'], errors='coerce', format='%Y-%m-%d
%H:%M:%S%z')

# Extract date and time components
df_twitter['Date'] = df_twitter['date'].dt.date
df_twitter['Time'] = df_twitter['date'].dt.time

# Drop the original 'date' column if not needed anymore
df_twitter.drop(columns=['date'], inplace=True)

# Print the resulting DataFrame
print(df_twitter[['Date', 'Time', 'id', 'content', 'username', 'like_count', 'retweet_count']])

```

```
# Analyzing Tweet Counts Over Different Months
```

```
# Convert 'Date' column to datetime format
```

```
df_twitter['Date'] = pd.to_datetime(df_twitter['Date'])
```

```
# Extract month and year from the 'Date' column
```

```
df_twitter['Month'] = df_twitter['Date'].dt.strftime('%B, %Y')
```

```
# Count by Month
```

```
fig, ax = plt.subplots(1, 1, figsize=(8, 6))
```

```
ax = sns.countplot(x=df_twitter['Month'], order=['January, 2023', 'February, 2023', 'March, 2023'],  
palette='Pastel1')
```

```
for p in ax.patches:
```

```
    ax.text(p.get_x() + p.get_width() / 2, p.get_height() + 1000, p.get_height(),  
           ha='center', va='center')
```

```
ax.set_title('Count of Tweets about ChatGPT by Months')
```

```
plt.xticks(rotation=45)
```

```
plt.tight_layout()
```

```
plt.show()
```

```
times_jan_feb = round(df_twitter['Month'].value_counts()['February, 2023'] /  
df_twitter['Month'].value_counts()['January, 2023'], 2)
```

```
times_feb_mar = round(df_twitter['Month'].value_counts()['March, 2023'] /  
df_twitter['Month'].value_counts()['February, 2023'], 2)
```

```
print(f'The tweets about ChatGPT in February, 2023 is {times_jan_feb} times of the tweets amount in  
January, 2023.')
```

```
print(f'The tweets about ChatGPT in March, 2023 is {times_feb_mar} times of the tweets amount in  
February, 2023.')
```

```
# Analyzing Weekly Tweet Counts and Activity Trends
```

```
# Convert 'Date' column to datetime format
```

```

df_twitter['Date'] = pd.to_datetime(df_twitter['Date'])

# Extract week number and year from the 'Date' column
df_twitter['Week'] = df_twitter['Date'].dt.strftime('%U, %Y')

# Filter data for the specified date range
start_date = '2023-01-04'
end_date = '2023-03-29'
filtered_tweets = df_twitter[(df_twitter['Date'] >= start_date) & (df_twitter['Date'] <= end_date)]

# Count by Week
fig, ax = plt.subplots(1, 1, figsize=(12, 6))
ax = sns.countplot(x=filtered_tweets['Week'], palette='PuBu')
ax.set_xticklabels([f'{i+1}st week' if i == 0 else f'{i+1}nd week' if i == 1 else f'{i+1}rd week' if i == 2 else
f'{i+1}th week' for i in range(len(filtered_tweets['Week'].unique()))], rotation=45)
ax.set_title('Count of Tweets by Weeks (January 2023 to March 2023)')
ax.set_xlabel('Week')
ax.set_ylabel('Tweet Count')
for p in ax.patches:
    ax.text(p.get_x() + p.get_width() / 2, p.get_height() + 1000, p.get_height(),
            ha='center', va='center', size=10)
plt.tight_layout()
plt.show()

# Find the week numbers for the specific dates
week_1st_jan = df_twitter[df_twitter['Date'] == '2023-01-04']['Week'].values[0]
week_4th_jan = df_twitter[df_twitter['Date'] == '2023-01-25']['Week'].values[0]
week_8th_feb = df_twitter[df_twitter['Date'] == '2023-02-22']['Week'].values[0]
week_last_mar = df_twitter[df_twitter['Date'] == '2023-03-29']['Week'].values[0]

# Calculate the ratio of tweet counts between specific weeks

```

```

times_1st_jan_4th_jan = round(df_twitter['Week'].value_counts()[week_1st_jan] /
df_twitter['Week'].value_counts()[week_4th_jan], 2)

times_4th_jan_8th_feb = round(df_twitter['Week'].value_counts()[week_4th_jan] /
df_twitter['Week'].value_counts()[week_8th_feb], 2)

times_8th_feb_last_mar = round(df_twitter['Week'].value_counts()[week_8th_feb] /
df_twitter['Week'].value_counts()[week_last_mar], 2)

print(f"The tweet activity from the 1st week to the 4th week increased by {times_1st_jan_4th_jan}
times.")

print(f"The tweet activity from the 4th week to the 8th week increased by {times_4th_jan_8th_feb}
times.")

print(f"The tweet activity from the 8th week to the last week increased by {times_8th_feb_last_mar}
times.")

#Distribution of Average Daily Tweets Across Months

# Extract month and year from the 'Date' column
df_twitter['Month'] = df_twitter['Date'].dt.strftime('%B')
df_twitter['Year'] = df_twitter['Date'].dt.year

# Filter data for January, February, and March
jan_tweets = df_twitter[(df_twitter['Month'] == 'January') & (df_twitter['Year'] == 2023)]
feb_tweets = df_twitter[(df_twitter['Month'] == 'February') & (df_twitter['Year'] == 2023)]
mar_tweets = df_twitter[(df_twitter['Month'] == 'March') & (df_twitter['Year'] == 2023)]

# Calculate average daily tweets for each month
avg_jan = jan_tweets.shape[0] / jan_tweets['Date'].dt.days_in_month.iloc[0]
avg_feb = feb_tweets.shape[0] / feb_tweets['Date'].dt.days_in_month.iloc[0]
avg_mar = mar_tweets.shape[0] / mar_tweets['Date'].dt.days_in_month.iloc[0]

print(f'Average daily tweets in January, 2023: {avg_jan:.2f}')
print(f'Average daily tweets in February, 2023: {avg_feb:.2f}')
print(f'Average daily tweets in March, 2023: {avg_mar:.2f}')

```

```

# Create a pie chart

plt.figure(figsize=(8, 8))

avg_daily_tweets = {
    'January': avg_jan,
    'February': avg_feb,
    'March': avg_mar
}

custom_colors = ['#66c2a5', '#fc8d62', '#8da0cb']

plt.pie(avg_daily_tweets.values(), labels=avg_daily_tweets.keys(), autopct='%1.1f%%',
startangle=140, colors=custom_colors)

plt.title('Distribution of Average Daily Tweets Across Months')

plt.axis('equal') # Equal aspect ratio ensures that the pie is drawn as a circle.

plt.show()

```

#Text Preprocessing

Define a function to clean and preprocess text

```

def preprocess_text(text):
    # Remove URLs using a more comprehensive regular expression
    text = re.sub(r'http[s]?://(?:[a-zA-Z]|[0-9]|[$-_@.&+]|[*\\(\)\.,]|(?:%[0-9a-fA-F][0-9a-fA-F]))+', '',
text)

    text = re.sub(r'@[A-Za-z0-9]+', '', text) # Remove mentions
    text = re.sub(r'#', '', text) # Remove hashtags
    text = re.sub(r'^A-Za-z\s', '', text) # Remove special characters
    return text

```

Apply preprocessing to the 'content' column

```

df_twitter['cleaned_content'] = df_twitter['content'].apply(preprocess_text)

```

```

def preprocess_tokens(text):
    tokens = word_tokenize(text.lower())

    stop_words = set(stopwords.words('english'))

```



```

filtered_tokens = [token for token in tokens if token not in stop_words]
stemmer = PorterStemmer()
lemmatizer = WordNetLemmatizer()
stemmed_tokens = [stemmer.stem(token) for token in filtered_tokens]
lemmatized_tokens = [lemmatizer.lemmatize(token) for token in filtered_tokens]
return lemmatized_tokens

df_twitter['preprocessed_tokens'] = df_twitter['cleaned_content'].apply(preprocess_tokens)

# Display the DataFrame with cleaned content and preprocessed tokens
print("\nDataFrame with Preprocessed Data:")
print(df_twitter[['content', 'cleaned_content', 'preprocessed_tokens']])

# Create a SentimentIntensityAnalyzer object
sia = SentimentIntensityAnalyzer()

# Define a function to get sentiment scores for each text
def get_sentiment_scores(text):
    sentiment_score = sia.polarity_scores(text)
    return sentiment_score

# Apply the function to each text in the 'content' column
df_twitter['sentiment_scores'] = df_twitter['cleaned_content'].apply(get_sentiment_scores)

# Extract the compound sentiment score and assign sentiment labels
df_twitter['compound_score'] = df_twitter['sentiment_scores'].apply(lambda x: x['compound'])

# Assign sentiment labels based on the compound score
df_twitter['sentiment_label'] = df_twitter['compound_score'].apply(
    lambda x: 'Positive' if x >= 0.05 else 'Negative' if x <= -0.05 else 'Neutral'

```

```
)
```

```
# Print the resulting DataFrame with sentiment labels
```

```
# Define a function to get sentiment scores using TextBlob
```

```
def get_sentiment_textblob(text):
```

```
    analysis = TextBlob(text)
```

```
    polarity = analysis.sentiment.polarity
```

```
    return polarity
```

```
# Create a SentimentIntensityAnalyzer object for VADER sentiment analysis
```

```
sia = SentimentIntensityAnalyzer()
```

```
# Define a function to get sentiment scores using VADER
```

```
def get_sentiment_vader(text):
```

```
    sentiment_score = sia.polarity_scores(text)
```

```
    compound_score = sentiment_score['compound']
```

```
    return compound_score
```

```
# Apply sentiment scoring using TextBlob and VADER to the 'content' column
```

```
df_twitter['sentiment_score_textblob'] =
```

```
df_twitter['cleaned_content'].apply(get_sentiment_textblob)
```

```
df_twitter['sentiment_score_vader'] = df_twitter['cleaned_content'].apply(get_sentiment_vader)
```

```
# Assign sentiment labels based on the sentiment scores
```

```
def assign_sentiment_label(score):
```

```
    if score > 0:
```

```
        return 'Positive'
```

```
    elif score < 0:
```

```
        return 'Negative'
```

```
    else:
```

```

        return 'Neutral'

df_twitter['sentiment_label_textblob'] =
df_twitter['sentiment_score_textblob'].apply(assign_sentiment_label)

df_twitter['sentiment_label_vader'] =
df_twitter['sentiment_score_vader'].apply(assign_sentiment_label)

# Define a function to categorize sentiment labels
def categorize_sentiment(polarity):
    if polarity > 0:
        return 'Positive'
    elif polarity < 0:
        return 'Negative'
    else:
        return 'Neutral'

# Apply the categorize_sentiment function to the 'sentiment_score_textblob' column
df_twitter['sentiment_category_textblob'] =
df_twitter['sentiment_score_textblob'].apply(categorize_sentiment)

# Apply the categorize_sentiment function to the 'sentiment_score_vader' column
df_twitter['sentiment_category_vader'] =
df_twitter['sentiment_score_vader'].apply(categorize_sentiment)

# Print the resulting DataFrame with sentiment labels and categories
print(df_twitter[['cleaned_content', 'sentiment_label_textblob', 'sentiment_label_vader',
'sentiment_category_textblob', 'sentiment_category_vader']])

# Count the number of tweets in each sentiment category
sentiment_counts_textblob = df_twitter['sentiment_category_textblob'].value_counts()
sentiment_counts_vader = df_twitter['sentiment_category_vader'].value_counts()

```

```

# Define color palettes (you can choose different palettes)
textblob_colors = sns.color_palette("pastel")
vader_colors = sns.color_palette("Pastel1")

# Create subplots for sentiment distribution using TextBlob and VADER
fig, axes = plt.subplots(1, 2, figsize=(12, 6))

# Sentiment distribution using TextBlob with custom colors
sns.barplot(x=sentiment_counts_textblob.index, y=sentiment_counts_textblob.values, ax=axes[0],
palette=textblob_colors)
axes[0].set_title('Sentiment Distribution (TextBlob)')
axes[0].set_xlabel('Sentiment Category')
axes[0].set_ylabel('Number of Tweets')

# Sentiment distribution using VADER with custom colors
sns.barplot(x=sentiment_counts_vader.index, y=sentiment_counts_vader.values, ax=axes[1],
palette=vader_colors)
axes[1].set_title('Sentiment Distribution (VADER)')
axes[1].set_xlabel('Sentiment Category')
axes[1].set_ylabel('Number of Tweets')

# Annotate the bars with the count of each sentiment category
for ax in axes:
    for p in ax.patches:
        ax.annotate(f'{int(p.get_height())}', (p.get_x() + p.get_width() / 2., p.get_height()),
            ha='center', va='center', fontsize=12, color='black', xytext=(0, 5),
            textcoords='offset points')

plt.tight_layout()
plt.show()

# Print the number of positive, negative, and neutral tweets

```

```

print("Number of Positive Tweets (TextBlob):", sentiment_counts_textblob['Positive'])
print("Number of Negative Tweets (TextBlob):", sentiment_counts_textblob['Negative'])
print("Number of Neutral Tweets (TextBlob):", sentiment_counts_textblob['Neutral'])

```

```

print("Number of Positive Tweets (VADER):", sentiment_counts_vader['Positive'])
print("Number of Negative Tweets (VADER):", sentiment_counts_vader['Negative'])
print("Number of Neutral Tweets (VADER):", sentiment_counts_vader['Neutral'])

```

```

# Count the number of tweets in each sentiment category
sentiment_counts_textblob = df_twitter['sentiment_category_textblob'].value_counts()
sentiment_counts_vader = df_twitter['sentiment_category_vader'].value_counts()

```

```

# Define custom lighter shades of colors for the pie chart
colors_light = ['#b4e88b', '#9bc1d2', '#f8a5a0']

```

```

# Create a 3D pie chart for sentiment distribution using TextBlob
fig1 = go.Figure(data=[go.Pie(labels=sentiment_counts_textblob.index,
                               values=sentiment_counts_textblob.values,
                               marker=dict(colors=colors_light),
                               textinfo='percent+label', textposition='inside',
                               sort=False, pull=[0, 0.1, 0]))]
fig1.update_layout(title_text='Sentiment Analysis Result (TextBlob)',
                    scene=dict(aspectmode="cube"))

```

```

# Create a 3D pie chart for sentiment distribution using VADER
fig2 = go.Figure(data=[go.Pie(labels=sentiment_counts_vader.index,
                               values=sentiment_counts_vader.values,
                               marker=dict(colors=colors_light),
                               textinfo='percent+label', textposition='inside',
                               sort=False, pull=[0, 0.1, 0]))]
fig2.update_layout(title_text='Sentiment Analysis Result (VADER)',

```

```

scene=dict(aspectmode="cube"))

# Display both 3D pie charts
fig1.show()
fig2.show()

# Group data by date and sentiment label
df_grouped_date = df_twitter.groupby(by=["Date"], as_index=False).count()

#Group data by date and sentiment label
df_grouped_date_sentiment = df_twitter.groupby(by=["Date", "sentiment_label_textblob"],
as_index=False).count()

# Group data by date and sentiment label (TextBlob)
df_grouped_date_sentiment_textblob = df_twitter.groupby(by=["Date",
"sentiment_label_textblob"], as_index=False).count()

df_grouped_date_sentiment_textblob =
df_grouped_date_sentiment_textblob.rename(columns={'content': 'tweet_count'})

# Get top 2 dates with the highest tweet count (TextBlob)
max_2_textblob = df_grouped_date_sentiment_textblob.groupby('Date',
as_index=False).sum().sort_values(by="tweet_count", ascending=False).iloc[:2]

# Create a single subplot for TextBlob
plt.figure(figsize=(12, 6))

# Choose a colormap for TextBlob sentiment trends
colormap_textblob = 'autumn_r'

# Plot Overall Trend in Tweet Count (TextBlob)
plt.plot(df_grouped_date["Date"], df_grouped_date["content"], label='Overall Tweet Count',
color='tab:red', linewidth=2)

```

```

plt.scatter(max_2_textblob["Date"], max_2_textblob["tweet_count"], color='tab:red', marker='o',
label='Top 2 Dates (TextBlob)')

plt.xlabel("Date")
plt.ylabel("Tweet Count")
plt.legend(loc='upper left')

# Plot Sentiment Trend as a shadow (TextBlob)
sentiments_textblob = df_grouped_date_sentiment_textblob["sentiment_label_textblob"].unique()
color_map_textblob = plt.cm.get_cmap(colormap_textblob, len(sentiments_textblob)) # Colormap
for different sentiments
for i, sentiment in enumerate(sentiments_textblob):
    sentiment_data =
df_grouped_date_sentiment_textblob[df_grouped_date_sentiment_textblob["sentiment_label_text
blob"] == sentiment]

    plt.fill_between(sentiment_data["Date"], sentiment_data["tweet_count"],
df_grouped_date["content"], alpha=0.3, label=f'{sentiment} Sentiment (TextBlob)',
color=color_map_textblob(i))

# Title and layout for TextBlob
plt.title("Overall Sentiment Tweet Count Over Time - TextBlob")
plt.tight_layout()

# Enhancements
plt.grid(True, linestyle='--', alpha=0.5)
plt.xticks(rotation=45)
plt.legend()
plt.show()

# Group data by date and sentiment label (VADER)
df_grouped_date_sentiment_vader = df_twitter.groupby(by=["Date", "sentiment_label_vader"],
as_index=False).count()

df_grouped_date_sentiment_vader =
df_grouped_date_sentiment_vader.rename(columns={'content': 'tweet_count'})

```

```

# Get top 2 dates with the highest tweet count (VADER)

max_2_vader = df_grouped_date_sentiment_vader.groupby('Date',
as_index=False).sum().sort_values(by="tweet_count", ascending=False).iloc[:2]


# Create a single subplot for VADER

plt.figure(figsize=(12, 6))


# Choose a colormap for VADER sentiment trends

colormap_vader = 'winter_r'


# Plot Overall Trend in Tweet Count (VADER)

plt.plot(df_grouped_date["Date"], df_grouped_date["content"], label='Overall Tweet Count',
color='tab:blue', linewidth=2)

plt.scatter(max_2_vader["Date"], max_2_vader["tweet_count"], color='tab:blue', marker='x',
label='Top 2 Dates (VADER)')

plt.xlabel("Date")

plt.ylabel("Tweet Count")

plt.legend(loc='upper left')


# Plot Sentiment Trend as a shadow (VADER)

sentiments_vader = df_grouped_date_sentiment_vader["sentiment_label_vader"].unique()

color_map_vader = plt.cm.get_cmap(colormap_vader, len(sentiments_vader))

# Colormap for different sentiments

for i, sentiment in enumerate(sentiments_vader):

    sentiment_data =
df_grouped_date_sentiment_vader[df_grouped_date_sentiment_vader["sentiment_label_vader"]
== sentiment]

    plt.fill_between(sentiment_data["Date"], sentiment_data["tweet_count"],
df_grouped_date["content"], alpha=0.3, label=f'{sentiment} Sentiment (VADER)',
color=color_map_vader(i))


# Title and layout for VADER

plt.title("Overall Sentiment Tweet Count Over Time - Vader")

```



```
plt.tight_layout()
```

```
# Enhancements
```

```
plt.grid(True, linestyle='--', alpha=0.5)
```

```
plt.xticks(rotation=45)
```

```
plt.legend()
```

```
plt.show()
```

```
# Assuming you have the 'Date' column in datetime format
```

```
df_twitter['Month'] = df_twitter['Date'].dt.month_name()
```

```
df_twitter['Day_of_Week'] = df_twitter['Date'].dt.day_name()
```

```
df_twitter['Hour_of_Day'] = df_twitter['Date'].dt.hour
```

```
# Group data by month, day of the week, and hour of the day
```

```
sentiment_heatmap = df_twitter.groupby(['Month', 'Day_of_Week', 'Hour_of_Day',  
    'sentiment_label_textblob']).size().unstack(fill_value=0)
```

```
# Create a heatmap
```

```
plt.figure(figsize=(15, 10))
```

```
sns.heatmap(sentiment_heatmap, cmap='YlGnBu', annot=True, fmt='d', linewidths=.5)
```

```
# Get the y-axis labels (month and day)
```

```
y_labels = sentiment_heatmap.index.get_level_values('Month') + ' - ' +  
    sentiment_heatmap.index.get_level_values('Day_of_Week')
```

```
plt.title('Sentiment Distribution by Month, Day of the Week, and Hour of the Day')
```

```
plt.xlabel('Hour of the Day')
```

```
plt.ylabel('Month - Day of the Week')
```

```
plt.xticks(rotation=45)
```

```
plt.yticks(ticks=range(len(y_labels)), labels=y_labels) # Set custom y-axis labels
```

```

plt.tight_layout()

plt.show()


# Assuming you have the 'Date' column in datetime format
df_twitter['Month'] = df_twitter['Date'].dt.month_name()
df_twitter['Day_of_Week'] = df_twitter['Date'].dt.day_name()
df_twitter['Hour_of_Day'] = df_twitter['Date'].dt.hour


# Group data by month, day of the week, and hour of the day
sentiment_heatmap = df_twitter.groupby(['Month', 'Day_of_Week', 'Hour_of_Day',
'sentiment_label_vader']).size().unstack(fill_value=0)


# Create a heatmap
plt.figure(figsize=(15, 10))

sns.heatmap(sentiment_heatmap, cmap='YlGnBu', annot=True, fmt='d', linewidths=.5)


# Get the y-axis labels (month and day)
y_labels = sentiment_heatmap.index.get_level_values('Month') + ' - ' +
sentiment_heatmap.index.get_level_values('Day_of_Week')


plt.title('Sentiment Distribution by Month, Day of the Week, and Hour of the Day')
plt.xlabel('Hour of the Day')
plt.ylabel('Month - Day of the Week')
plt.xticks(rotation=45)
plt.yticks(ticks=range(len(y_labels)), labels=y_labels) # Set custom y-axis labels
plt.tight_layout()

plt.show()


from wordcloud import WordCloud

```

```

# Combine all tweet content into a single string
all_tweets_text = ''.join(df_twitter['cleaned_content'])

# Generate a word cloud
wordcloud = WordCloud(width=800, height=400, background_color='white',
colormap='viridis').generate(all_tweets_text)

# Create a figure and axis
plt.figure(figsize=(10, 6))
plt.imshow(wordcloud, interpolation='bilinear')
plt.title('Word Cloud of Most Frequent Words in Tweets')
plt.axis('off')
plt.show()

# Group the data by 'Month' and calculate the count of each sentiment category
textblob_sentiment_counts = df_twitter.groupby(['Month',
'sentiment_category_textblob']).size().unstack(fill_value=0)

vader_sentiment_counts = df_twitter.groupby(['Month',
'sentiment_category_vader']).size().unstack(fill_value=0)

# Plot sentiment distribution over time (by month) for TextBlob
plt.figure(figsize=(12, 6))
sns.set(style="whitegrid")
sns.lineplot(data=textblob_sentiment_counts, dashes=False)
plt.title('Sentiment Distribution Over Time (TextBlob)')
plt.xlabel('Month')
plt.ylabel('Count')
plt.legend(title='Sentiment', loc='upper right', labels=['Negative', 'Neutral', 'Positive'])
plt.xticks(rotation=45)
plt.tight_layout()

```

```
# Show the plot
```

```
plt.show()
```

```
# Plot sentiment distribution over time (by month) for VADER
```

```
plt.figure(figsize=(12, 6))
```

```
sns.set(style="whitegrid")
```

```
sns.lineplot(data=vader_sentiment_counts, dashes=False)
```

```
plt.title('Sentiment Distribution Over Time (VADER)')
```

```
plt.xlabel('Month')
```

```
plt.ylabel('Count')
```

```
plt.legend(title='Sentiment', loc='upper right', labels=['Negative', 'Neutral', 'Positive'])
```

```
plt.xticks(rotation=45)
```

```
plt.tight_layout()
```

```
# Show the plot
```

```
plt.show()
```

```
# Plot sentiment distribution by month
```

```
df_twitter['Month'] = df_twitter['Date'].dt.to_period('M')
```

```
plt.figure(figsize=(10, 6))
```

```
sns.set_style('whitegrid')
```

```
sns.countplot(data=df_twitter, x='Month', hue='sentiment_label', palette='pastel')
```

```
plt.title('Sentiment Distribution Over Time (By Month)')
```

```
plt.xticks(rotation=45)
```

```
plt.tight_layout()
```

```
plt.show()
```

```
# Plot sentiment distribution by week
```

```
df_twitter['Week'] = df_twitter['Date'].dt.to_period('W')
```

```

plt.figure(figsize=(10, 6))

sns.set_style('whitegrid')

sns.countplot(data=df_twitter, x='Week', hue='sentiment_label', palette='pastel')

plt.title('Sentiment Distribution Over Time (By Week)')

plt.xticks(rotation=45)

plt.tight_layout()

plt.show()


# Extract week and year from the 'Date' column

df_twitter['Week'] = df_twitter['Date'].dt.strftime('%U')

df_twitter['Year'] = df_twitter['Date'].dt.strftime('%Y')


# Filter data for the first and last week

first_week = df_twitter[(df_twitter['Year'] == '2023') & (df_twitter['Week'] ==
df_twitter[df_twitter['Year'] == '2023']['Week'].min())]

last_week = df_twitter[(df_twitter['Year'] == '2023') & (df_twitter['Week'] ==
df_twitter[df_twitter['Year'] == '2023']['Week'].max())]


# Calculate sentiment counts for both weeks and sentiment labels (TextBlob)

first_week_sentiments_textblob =
first_week['sentiment_label_textblob'].value_counts().reindex(df_twitter['sentiment_label_textblob']
].unique(), fill_value=0)

last_week_sentiments_textblob =
last_week['sentiment_label_textblob'].value_counts().reindex(df_twitter['sentiment_label_textblob']
].unique(), fill_value=0)


# Calculate sentiment counts for both weeks and sentiment labels (VADER)

first_week_sentiments_vader =
first_week['sentiment_label_vader'].value_counts().reindex(df_twitter['sentiment_label_vader'].unique(), fill_value=0)

last_week_sentiments_vader =
last_week['sentiment_label_vader'].value_counts().reindex(df_twitter['sentiment_label_vader'].unique(), fill_value=0)

```

```

# Calculate the differences between the first and last week sentiment counts
difference_textblob = last_week_sentiments_textblob - first_week_sentiments_textblob
difference_vader = last_week_sentiments_vader - first_week_sentiments_vader


# Define the sentiment labels
sentiment_labels = df_twitter['sentiment_label_textblob'].unique()


# Set the width of the bars
bar_width = 0.35


# Create an array of indices for the sentiment labels
indices = np.arange(len(sentiment_labels))


# Create subplots for TextBlob and VADER sentiment comparisons
fig, axes = plt.subplots(2, 2, figsize=(14, 10))


# Plot TextBlob sentiment distribution comparison
axes[0, 0].bar(indices - bar_width/2, first_week_sentiments_textblob, bar_width, label='First Week',
alpha=0.7)
axes[0, 0].bar(indices + bar_width/2, last_week_sentiments_textblob, bar_width, label='Last Week',
alpha=0.7)
axes[0, 0].set_xlabel('Sentiment Label')
axes[0, 0].set_ylabel('Tweet Count')
axes[0, 0].set_title('Sentiment Distribution Comparison (TextBlob)')
axes[0, 0].set_xticks(indices)
axes[0, 0].set_xticklabels(sentiment_labels, rotation=45)
axes[0, 0].legend()


# Display the differences for TextBlob sentiment
for i, diff in enumerate(difference_textblob):

```

```
axes[0, 0].text(i, max(first_week_sentiments_textblob[i], last_week_sentiments_textblob[i]) + 10,
f"Diff: {diff}", ha='center', va='bottom', color='red')
```

```
# Plot VADER sentiment distribution comparison
```

```
axes[0, 1].bar(indices - bar_width/2, first_week_sentiments_vader, bar_width, label='First Week',
alpha=0.7)
```

```
axes[0, 1].bar(indices + bar_width/2, last_week_sentiments_vader, bar_width, label='Last Week',
alpha=0.7)
```

```
axes[0, 1].set_xlabel('Sentiment Label')
```

```
axes[0, 1].set_ylabel('Tweet Count')
```

```
axes[0, 1].set_title('Sentiment Distribution Comparison (VADER)')
```

```
axes[0, 1].set_xticks(indices)
```

```
axes[0, 1].set_xticklabels(sentiment_labels, rotation=45)
```

```
axes[0, 1].legend()
```

```
# Display the differences for VADER sentiment
```

```
for i, diff in enumerate(difference_vader):
```

```
    axes[0, 1].text(i, max(first_week_sentiments_vader[i], last_week_sentiments_vader[i]) + 10, f"Diff:
{diff}", ha='center', va='bottom', color='red')
```

```
# Hide unnecessary subplots
```

```
axes[1, 0].axis('off')
```

```
axes[1, 1].axis('off')
```

```
# Add a legend for both subplots
```

```
plt.tight_layout()
```

```
plt.show()
```

```
# Extract month and year from the 'Date' column
```

```
df_twitter['Year'] = df_twitter['Date'].dt.year
```

```
df_twitter['Month'] = df_twitter['Date'].dt.month
```

```

# Extract month and year from the 'Date' column
df_twitter['Year'] = df_twitter['Date'].dt.year
df_twitter['Month'] = df_twitter['Date'].dt.month

# Filter data for the first and last month
first_month = df_twitter[df_twitter['Year'] == 2023]['Month'].min()
last_month = df_twitter[df_twitter['Year'] == 2023]['Month'].max()
df_first_month = df_twitter[(df_twitter['Year'] == 2023) & (df_twitter['Month'] == first_month)]
df_last_month = df_twitter[(df_twitter['Year'] == 2023) & (df_twitter['Month'] == last_month)]

# Calculate sentiment percentages for both months and sentiment labels (TextBlob)
first_month_sentiments_textblob =
df_first_month['sentiment_label_textblob'].value_counts(normalize=True).reindex(df_twitter['sentiment_label_textblob'].unique(), fill_value=0)

last_month_sentiments_textblob =
df_last_month['sentiment_label_textblob'].value_counts(normalize=True).reindex(df_twitter['sentiment_label_textblob'].unique(), fill_value=0)

# Calculate sentiment percentages for both months and sentiment labels (VADER)
first_month_sentiments_vader =
df_first_month['sentiment_label_vader'].value_counts(normalize=True).reindex(df_twitter['sentiment_label_vader'].unique(), fill_value=0)

last_month_sentiments_vader =
df_last_month['sentiment_label_vader'].value_counts(normalize=True).reindex(df_twitter['sentiment_label_vader'].unique(), fill_value=0)

# Plot sentiment distribution for the first and last months using stacked bar charts with percentages
plt.figure(figsize=(14, 6))

Pale_Pink = (255/255, 182/255, 193/255)
lavender_color = (230/255, 230/255, 250/255)

bar_width = 0.35
index = range(len(df_twitter['sentiment_label_textblob'].unique()))

```



```

plt.subplot(1, 2, 1)

plt.bar(index, first_month_sentiments_textblob * 100, width=bar_width, label='First Month
(TextBlob)', color=lavender_color)

plt.bar(index, last_month_sentiments_textblob * 100, width=bar_width, label='Last Month
(TextBlob)', bottom=first_month_sentiments_textblob * 100, color=Pale_Pink)

plt.xlabel('Sentiment Label (TextBlob)')

plt.ylabel('Percentage')

plt.title(f'Percentage Distribution Comparison for 1st Month and last Month (TextBlob)')

plt.xticks(index, df_twitter['sentiment_label_textblob'].unique(), rotation=45)


# Display the percentages for both "first month" and "last month" TextBlob data
for i, (percentage1, percentage2) in enumerate(zip(first_month_sentiments_textblob * 100,
last_month_sentiments_textblob * 100)):

    plt.text(i, (percentage1 + percentage2) / 2, f"{percentage1:.2f}% / {percentage2:.2f}%", ha='center',
va='bottom')


plt.subplot(1, 2, 2)

plt.bar(index, first_month_sentiments_vader * 100, width=bar_width, label='First Month (VADER)',
color=lavender_color)

plt.bar(index, last_month_sentiments_vader * 100, width=bar_width, label='Last Month (VADER)',
bottom=first_month_sentiments_vader * 100, color=Pale_Pink)

plt.xlabel('Sentiment Label (VADER)')

plt.ylabel('Percentage')

plt.title(f'Percentage Distribution Comparison for 1st Month and last Month (VADER)')

plt.xticks(index, df_twitter['sentiment_label_vader'].unique(), rotation=45)


# Display the percentages for both "first month" and "last month" VADER data
for i, (percentage1, percentage2) in enumerate(zip(first_month_sentiments_vader * 100,
last_month_sentiments_vader * 100)):

    plt.text(i, (percentage1 + percentage2) / 2, f"{percentage1:.2f}% / {percentage2:.2f}%", ha='center',
va='bottom')


plt.tight_layout()

plt.legend()

```

```

plt.show()

# Filter negative reviews from both VADER and TextBlob results
negative_reviews_vader = df_twitter[df_twitter['sentiment_label_vader'] == 'Negative']
negative_reviews_textblob = df_twitter[df_twitter['sentiment_label_textblob'] == 'Negative']

# Group negative reviews by date (e.g., by week or month)
negative_reviews_vader_by_date =
negative_reviews_vader.groupby('Date').size().reset_index(name='NegativeCount_VADER')

negative_reviews_textblob_by_date =
negative_reviews_textblob.groupby('Date').size().reset_index(name='NegativeCount_TextBlob')

# Merge the two DataFrames based on the 'Date' column
negative_reviews_combined = pd.merge(negative_reviews_vader_by_date,
negative_reviews_textblob_by_date, on='Date', how='outer')

# Fill missing values with 0
negative_reviews_combined = negative_reviews_combined.fillna(0)

# Plot negative review counts over time using a line plot
plt.figure(figsize=(12, 6))

plt.plot(negative_reviews_combined['Date'], negative_reviews_combined['NegativeCount_VADER'],
label='VADER Negative Reviews')

plt.plot(negative_reviews_combined['Date'],
negative_reviews_combined['NegativeCount_TextBlob'], label='TextBlob Negative Reviews')

plt.title('Negative Review Trends Over Time')
plt.xlabel('Date')
plt.ylabel('Number of Negative Reviews')
plt.legend()
plt.xticks(rotation=45)
plt.tight_layout()

```

```

# Filter out rows with non-numeric sentiment values
df_twitter = df_twitter[df_twitter['sentiment_label_textblob'].isin(['Positive', 'Neutral', 'Negative'])]
df_twitter = df_twitter[df_twitter['sentiment_label_vader'].isin(['Positive', 'Neutral', 'Negative'])]

# Remove rows with non-numeric values in 'like_count' and 'retweet_count'
df_twitter = df_twitter[pd.to_numeric(df_twitter['like_count'], errors='coerce').notnull()]
df_twitter = df_twitter[pd.to_numeric(df_twitter['retweet_count'], errors='coerce').notnull()]

# Convert 'like_count' and 'retweet_count' to numeric
df_twitter['like_count'] = df_twitter['like_count'].astype(float)
df_twitter['retweet_count'] = df_twitter['retweet_count'].astype(float)

# Map sentiment categories to numeric values
sentiment_mapping = {'Positive': 1, 'Neutral': 0, 'Negative': -1}
df_twitter['textblob_sentiment_numeric'] =
df_twitter['sentiment_label_textblob'].map(sentiment_mapping)
df_twitter['vader_sentiment_numeric'] =
df_twitter['sentiment_label_vader'].map(sentiment_mapping)

# Calculate the correlation matrix
correlation_matrix = df_twitter[['like_count', 'retweet_count', 'textblob_sentiment_numeric',
'vader_sentiment_numeric']].corr()

# Create a heatmap to visualize the correlation
plt.figure(figsize=(10, 6))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', linewidths=0.5)
plt.title('Correlation Analysis: Sentiment vs. User Engagement')
plt.show()

#polarity
# Define a function to get sentiment polarity
def get_sentiment_polarity(text):

```

```

analysis = TextBlob(text)

return analysis.sentiment.polarity


# Apply the function to your DataFrame's text column
df_twitter['sentiment_polarity'] = df_twitter['cleaned_content'].apply(get_sentiment_polarity)

# Define a function to get sentiment polarity using VADER
def get_sentiment_polarity_vader(text):
    sentiment = sia.polarity_scores(text)
    return sentiment['compound']


# Apply the function to your DataFrame's text column
df_twitter['sentiment_polarity_vader'] =
df_twitter['cleaned_content'].apply(get_sentiment_polarity_vader)


# Set the style for the plots
sns.set_style("whitegrid")


# Create subplots
plt.figure(figsize=(12, 6))


# Plot the histogram for TextBlob sentiment polarity
plt.subplot(1, 2, 1)

sns.histplot(df_twitter['sentiment_polarity'], bins=20, color='blue', kde=True)

plt.title('TextBlob Sentiment Polarity Distribution')

plt.xlabel('Sentiment Polarity')

plt.ylabel('Frequency')


# Plot the histogram for VADER sentiment polarity
plt.subplot(1, 2, 2)

sns.histplot(df_twitter['sentiment_polarity_vader'], bins=20, color='green', kde=True)

plt.title('VADER Sentiment Polarity Distribution')

```

```
plt.xlabel('Sentiment Polarity')
plt.ylabel('Frequency')

# Add a title to the overall figure
plt.suptitle('Sentiment Polarity Analysis', fontsize=16)

# Adjust spacing between subplots
plt.tight_layout()

# Show the overlaid histograms
plt.show()

# Split the data into training and testing sets
X = df_twitter['cleaned_content']
y = df_twitter['sentiment_label']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Vectorize the text data using CountVectorizer
vectorizer = CountVectorizer()
X_train_vectorized = vectorizer.fit_transform(X_train)
X_test_vectorized = vectorizer.transform(X_test)

# Train a Logistic Regression model
logistic_regression_model = LogisticRegression()
logistic_regression_model.fit(X_train_vectorized, y_train)

# Predict sentiment labels on the test data
y_pred = logistic_regression_model.predict(X_test_vectorized)

# Evaluate the model
```

```

accuracy = accuracy_score(y_test, y_pred)
report = classification_report(y_test, y_pred)

# Print the evaluation results
print("Logistic Regression Model:")
print(f"Accuracy: {accuracy:.2f}")
print("Classification Report:\n", report)

# Train a Naive Bayes (MultinomialNB) model
naive_bayes_model = MultinomialNB()
naive_bayes_model.fit(X_train_vectorized, y_train)

# Predict sentiment labels on the test data
y_pred_nb = naive_bayes_model.predict(X_test_vectorized)

# Evaluate the Naive Bayes model
accuracy_nb = accuracy_score(y_test, y_pred_nb)
report_nb = classification_report(y_test, y_pred_nb)

# Print the evaluation results for Naive Bayes
print("\nNaive Bayes Model:")
print(f"Accuracy: {accuracy_nb:.2f}")
print("Classification Report:\n", report_nb)

# Calculate and print evaluation metrics for Logistic Regression
print("Logistic Regression Evaluation:")
print(f"Accuracy: {accuracy_score(y_test, y_pred):.2f}")
print(f"Precision: {precision_score(y_test, y_pred, average='weighted'):.2f}")
print(f"Recall: {recall_score(y_test, y_pred, average='weighted'):.2f}")

```

```

print(f"F1-Score: {f1_score(y_test, y_pred, average='weighted'):.2f}")

# Calculate and print evaluation metrics for Naive Bayes
print("\nNaive Bayes Evaluation:")
print(f"Accuracy: {accuracy_score(y_test, y_pred_nb):.2f}")
print(f"Precision: {precision_score(y_test, y_pred_nb, average='weighted'):.2f}")
print(f"Recall: {recall_score(y_test, y_pred_nb, average='weighted'):.2f}")
print(f"F1-Score: {f1_score(y_test, y_pred_nb, average='weighted'):.2f}")

# Calculate accuracy
accuracy_textblob = accuracy_score(df_twitter['sentiment_label'],
df_twitter['sentiment_category_textblob'])

accuracy_vader = accuracy_score(df_twitter['sentiment_label'],
df_twitter['sentiment_category_vader'])

# Calculate precision, recall, and F1-score
precision_textblob = precision_score(df_twitter['sentiment_label'],
df_twitter['sentiment_category_textblob'], average='weighted')

recall_textblob = recall_score(df_twitter['sentiment_label'],
df_twitter['sentiment_category_textblob'], average='weighted')

f1_score_textblob = f1_score(df_twitter['sentiment_label'],
df_twitter['sentiment_category_textblob'], average='weighted')

precision_vader = precision_score(df_twitter['sentiment_label'],
df_twitter['sentiment_category_vader'], average='weighted')

recall_vader = recall_score(df_twitter['sentiment_label'], df_twitter['sentiment_category_vader'],
average='weighted')

f1_score_vader = f1_score(df_twitter['sentiment_label'], df_twitter['sentiment_category_vader'],
average='weighted')

# Print the metrics
print("TextBlob Sentiment Analysis Metrics:")
print(f"Accuracy: {accuracy_textblob}")
print(f"Precision: {precision_textblob}")

```

```

print(f"Recall: {recall_textblob}")
print(f"F1-score: {f1_score_textblob}")

print("\nVADER Sentiment Analysis Metrics:")
print(f"Accuracy: {accuracy_vader}")
print(f"Precision: {precision_vader}")
print(f"Recall: {recall_vader}")
print(f"F1-score: {f1_score_vader}")

# Combine all content into a single string
all_content = " ".join(df_twitter['cleaned_content'])

# Tokenize the content (split into words)
words = all_content.split()

# Count the occurrences of each word
word_counts = Counter(words)

# Find the top most occurring words
top_words = word_counts.most_common(30)

# Print the top words and their frequencies
for word, frequency in top_words:
    print(f"{word}: {frequency}")

# List of words to remove
words_to_remove = [
    'the', 'to', '#ChatGPT', 'and', 'a', 'of', 'is', 'in', 'for', 'I', 'it', 'with', 'artificialintelligence', 'google',
    'on', 'Chat', 'you', 'GPT', '#AI', 'AI', 'that', 'ChatGPT', 'be', '#chatgpt', 'chatbot'
    'chat', 'this', 'can', 'are', 'about', 'will', 'by', 'your', 'gpt', 'gpt4', 'chatgpt', 'https', 'ai', 'openai'
]

```



```

def preprocess_text(text):

    # Remove URLs
    text = re.sub(r'http\S+', '', text)

    # Remove mentions (e.g., @user)
    text = re.sub(r'@\w+', '', text)

    # Convert text to lowercase
    text = text.lower()

    # Remove punctuation
    text = text.translate(str.maketrans("", "", string.punctuation))

    # Remove hashtags (e.g., #example)
    text = re.sub(r'#\w+', '', text)

    # Remove specific words
    words = text.split()
    cleaned_words = [word for word in words if word.lower() not in words_to_remove]

    return ' '.join(cleaned_words)

# Apply the custom function to your text column
df_twitter['cleaned_content'] = df_twitter['content'].apply(preprocess_text)

# Build bigram and trigram models
bigram = Phrases(df_twitter['preprocessed_tokens'], min_count=100, threshold=100)
bigram_mod = Phraser(bigram)

trigram = Phrases(bigram_mod[df_twitter['preprocessed_tokens']], min_count=100, threshold=100)

```

```

trigram_mod = Phraser(trigram)

# Apply bigram and trigram models to your preprocessed tokens
df_twitter['bigram_tokens'] = df_twitter['preprocessed_tokens'].apply(lambda x: bigram_mod[x])
df_twitter['trigram_tokens'] = df_twitter['bigram_tokens'].apply(lambda x: trigram_mod[x])

# Define a list of your specific topics
specific_topics = ["student", "students", "education", "job", "employment", "risk"]

# Convert your preprocessed text data into a list of strings
documents = df_twitter['preprocessed_tokens'].apply(lambda x: ' '.join(x))

# Initialize the CountVectorizer
vectorizer = CountVectorizer(max_df=0.95, min_df=2, max_features=1000, stop_words='english')

# Fit and transform your text data using CountVectorizer
X = vectorizer.fit_transform(documents)

# Initialize LDA with the desired number of topics
num_topics = 10 # You can adjust the number of topics
lda_model = LatentDirichletAllocation(n_components=num_topics, random_state=42)

# Fit the LDA model to your data
lda_topic_matrix = lda_model.fit_transform(X)

# Assign topics to tweets
df_twitter['dominant_topic'] = lda_topic_matrix.argmax(axis=1)

# Print the distribution of topics
topic_counts = df_twitter['dominant_topic'].value_counts().sort_index()
print("Distribution of Topics:")

```

```

print(topic_counts.head(5))

# Get the feature names (words) from the vectorizer
feature_names = vectorizer.get_feature_names_out()

# Get the top words for each topic
def get_top_words_for_topic_with_removal(model, feature_names, n_words=10,
words_to_remove=None):
    topics = []
    for topic_idx, topic in enumerate(model.components_):
        top_words_idx = topic.argsort()[-n_words:][::-1]
        top_words = [feature_names[i] for i in top_words_idx if feature_names[i] not in
words_to_remove]
        topics.append((topic_idx, top_words))
    return topics

# Get the top words for each topic while removing specified words
topics = get_top_words_for_topic_with_removal(lda_model, feature_names,
words_to_remove=words_to_remove)

top_5_topics = topics[:5]

# Print the top 5 topics and their top words
for topic_idx, top_words in top_5_topics:
    print(f"Topic {topic_idx + 1}: {' '.join(top_words)}")

#assign topics to each document in your DataFrame
def assign_topics_to_documents(model, X):
    # Get the topic probabilities for each document
    topic_probabilities = model.transform(X)

    # Find the dominant topic for each document

```

```

dominant_topics = [np.argmax(topic_probs) for topic_probs in topic_probabilities]

return dominant_topics

# Assign topics to documents
df_twitter['dominant_topic'] = assign_topics_to_documents(lda_model, X)

# Get the top 5 topics by sorting and selecting the first 5 rows
top_5_topics = topic_counts.sort_values(ascending=False).head(5)

# Filter the DataFrame to include only the top 5 topics
df_top_5_topics = df_twitter[df_twitter['dominant_topic'].isin(top_5_topics.index)]

# Define custom labels for the x-axis based on topic names
custom_labels = [
    "Topic 1", "Topic 2", "Topic 3", "Topic 4", "Topic 5"
]

# Create a bar plot for the top 5 topics with custom labels
plt.figure(figsize=(10, 6))
sns.countplot(data=df_top_5_topics, x='dominant_topic', palette='crest')
plt.title('Distribution of Top 5 Dominant Topics (LDA)')
plt.xlabel('Dominant Topic')
plt.ylabel('Tweet Count')
plt.xticks(range(len(custom_labels)), custom_labels, rotation=0)
plt.tight_layout()
plt.show()

# Define the top 5 topic headings
topic_headings = [
    "Topic 1: AI in Education and Job Market",

```

```

"Topic 2: ChatGPT in SEO and Content Marketing",
"Topic 3: ChatGPT Language Model and OpenAI",
"Topic 4: AI and Chatbots in Technology Industry",
"Topic 5: Future of AI and Human Technology"
]

# Get the top words for each topic

topics = get_top_words_for_topic_with_removal(lda_model, feature_names,
words_to_remove=words_to_remove)

# Print the top 5 topics and their headings side by side
for topic_idx, (topic_words, heading) in enumerate(zip(topics[:5], topic_headings)):
    top_words = ' '.join(topic_words[1]) # Join the top words into a single string
    print(f"{heading}\n{top_words}\n")

# Count the number of tweets in each of the top 5 topics
topic_counts = df_top_5_topics['dominant_topic'].value_counts()

# Create a pie chart
plt.figure(figsize=(8, 8))
patches, texts, autotexts = plt.pie(
    topic_counts,
    labels=custom_labels,
    autopct='%1.1f%%',
    startangle=140,
    colors=sns.color_palette('crest', len(custom_labels))
)

# Add topic names as annotations
for i, text in enumerate(texts):
    text.set_text(f"{custom_labels[i]}\n({topic_headings[i].split(':')[1].strip()})")

```

```

plt.title('Distribution of Top 5 Dominant Topics (LDA)')

plt.axis('equal')


# Show the pie chart

plt.show()


# Calculate the prevalence of the top 5 topics

top_5_topic_prevalence =
df_twitter['dominant_topic'].value_counts(normalize=True).sort_index().head(5)


# Print the topic prevalence

print("Topic Prevalence for the Top 5 Topics:")

print(top_5_topic_prevalence)


# Create a bar chart to visualize topic prevalence

plt.figure(figsize=(10, 6))

plt.bar(top_5_topic_prevalence.index, top_5_topic_prevalence.values)

plt.xlabel('Topic')

plt.ylabel('Prevalence')

plt.title('Topic Prevalence for the Top 5 Topics')

plt.xticks(top_5_topic_prevalence.index)

plt.show()


# Group the DataFrame by 'dominant_topic'

topic_groups = df_twitter.groupby('dominant_topic')


# Initialize empty lists to store topic and average sentiment scores

topics = []

average_sentiments = []

```

```

# Iterate through each topic group
for topic, group in topic_groups:

    # Calculate the average compound sentiment score for the group
    average_score = group['sentiment_score_vader'].mean()

    # Append the topic and average sentiment score to the lists
    topics.append(topic)
    average_sentiments.append(average_score)

# Create a new DataFrame to store the results
topic_sentiment_df = pd.DataFrame({'Topic': topics, 'Average Sentiment Score':
average_sentiments})

# Print or visualize the topic sentiment DataFrame
topic_sentiment_df.head(5)

# Define a function to calculate the percentage of each sentiment category
def calculate_sentiment_percentage(group):
    total_count = len(group)
    positive_count = (group['sentiment_category_vader'] == 'Positive').sum()
    negative_count = (group['sentiment_category_vader'] == 'Negative').sum()
    neutral_count = (group['sentiment_category_vader'] == 'Neutral').sum()

    positive_percentage = (positive_count / total_count) * 100
    negative_percentage = (negative_count / total_count) * 100
    neutral_percentage = (neutral_count / total_count) * 100

    return {
        'Positive Percentage': positive_percentage,
        'Negative Percentage': negative_percentage,
        'Neutral Percentage': neutral_percentage,
    }

```

```
}
```

```
# Group the DataFrame by 'dominant_topic' and apply the function to calculate percentages
```

```
sentiment_percentages =  
df_twitter.groupby('dominant_topic').apply(calculate_sentiment_percentage)
```

```
# Reset the index to have 'dominant_topic' as a regular column
```

```
sentiment_percentages = sentiment_percentages.reset_index()
```

```
# Create separate columns for Positive, Negative, and Neutral Percentages
```

```
sentiment_percentages['Positive'] = sentiment_percentages[0].apply(lambda x: x['Positive  
Percentage'])
```

```
sentiment_percentages['Negative'] = sentiment_percentages[0].apply(lambda x: x['Negative  
Percentage'])
```

```
sentiment_percentages['Neutral'] = sentiment_percentages[0].apply(lambda x: x['Neutral  
Percentage'])
```

```
# Drop the original column containing dictionaries
```

```
sentiment_percentages.drop(columns=[0], inplace=True)
```

```
# Print or visualize the sentiment percentages DataFrame
```

```
print(sentiment_percentages.head(5))
```

```
# Extract data for plotting
```

```
topics = sentiment_percentages['dominant_topic'][:5]
```

```
positive_percentage = sentiment_percentages['Positive'][:5]
```

```
negative_percentage = sentiment_percentages['Negative'][:5]
```

```
neutral_percentage = sentiment_percentages['Neutral'][:5]
```

```
# Set the width of the bars
```

```
bar_width = 0.2
```



```

# Create an array of indices for the topics
indices = np.arange(len(topics))

# Define lighter colors
light_green = '#90EE90'
light_red = '#FFC0CB'
light_blue = '#ADD8E6'

# Create the grouped bar chart with lighter colors
plt.figure(figsize=(12, 6))

plt.bar(indices - bar_width, positive_percentage, bar_width, label='Positive', color=light_green)
plt.bar(indices, negative_percentage, bar_width, label='Negative', color=light_red)
plt.bar(indices + bar_width, neutral_percentage, bar_width, label='Neutral', color=light_blue)

# Customize the chart
plt.xlabel('Topics')
plt.ylabel('Percentage')
plt.title('Sentiment Percentage by Topic for the First 5 Topics')
plt.xticks(indices, topics, rotation=45, ha='right')
plt.legend()

# Annotate percentages on the bars
for i, ind in enumerate(indices):
    plt.annotate(f'{positive_percentage[i]:.2f}%', (ind - bar_width / 2, positive_percentage[i] + 1))
    plt.annotate(f'{negative_percentage[i]:.2f}%', (ind, negative_percentage[i] + 1))
    plt.annotate(f'{neutral_percentage[i]:.2f}%', (ind + bar_width / 2, neutral_percentage[i] + 1))

# Show the plot
plt.tight_layout()
plt.show()

```

