

**Department of Electrical and Computer Engineering**

**North South University**

**CSE499A**

**Sentiment Analysis using Twitter 2023 Dataset**

|  |  |
| --- | --- |
| **Jannatun Nayem** | **1821896642** |

**Faculty Advisor:**

**DR. MOHAMMAD ASHRAFUZZAMAN KHAN**

**Assistant Professor**

**ECE Department**

**Summer, 2023**

**Abstract**

This sentiment analysis project focused on extracting insights from Twitter data using advanced models. The primary objectives included evaluating the accuracy of sentiment analysis, particularly with the BERT-based model, and understanding the broader implications of sentiment analysis in the context of social media. The project achieved a remarkable 98% accuracy, with a validation accuracy of 81%, showcasing the effectiveness of the chosen model. The findings underscore the significance of sentiment analysis in extracting valuable insights from social media platforms like Twitter, with applications in decision-making processes for businesses, governments, and researchers.

**Chapter 1: Introduction**

In the contemporary digital landscape, the role of social media platforms, particularly Twitter, has undergone a profound transformation, becoming not only a medium for individual expression but also a vast reservoir of collective sentiments and opinions. Recognizing the evolving nature of discourse on these platforms and the limitations of traditional sentiment analysis methodologies, this project embarked on a comprehensive exploration, utilizing the BERT-based model to unravel the intricacies of sentiments embedded in real-time Twitter data. The primary motivation behind this endeavor lies in the recognition that the brevity and rapidity of communication on Twitter demand advanced models to accurately capture the diverse range of emotions expressed. The purpose of this project is twofold: to rigorously assess the accuracy of sentiment analysis using the BERT model and to contribute nuanced insights that advance our broader understanding of sentiment analysis methodologies, particularly within the dynamic realm of social media.

Commencing with an exploration of the motivations and objectives of the project, the introduction sets the stage by acknowledging the transformative influence of social media platforms in shaping contemporary discourse. It underscores the challenges posed by the sheer volume and velocity of data generated on platforms like Twitter, where sentiments are compressed into 280 characters, making traditional sentiment analysis tools less effective. The introduction establishes the need for advanced models and positions the project as a response to this critical gap, emphasizing the relevance of the chosen BERT-based model in navigating the complexities of sentiment analysis.

The literature review serves as the intellectual foundation of the project, providing a panoramic survey of existing research in the field of sentiment analysis. Traditional approaches are scrutinized for their limitations, particularly in the context of social media. The review highlights the emergence of BERT as a revolutionary model capable of contextual understanding, a feature deemed essential for deciphering sentiment in the intricate tapestry of Twitter data. It meticulously positions the project within the evolving landscape of sentiment analysis, justifying the selection of the BERT-based model as an innovative and effective solution to the challenges posed by social media discourse.

The methodology section serves as a detailed roadmap, elucidating the systematic approach undertaken to navigate the complexities of sentiment analysis on Twitter. The utilization of the Twitter API for data extraction is detailed, emphasizing the relevance of real-time, user-generated content. Preprocessing steps, including noise removal, handling emoticons and hashtags, and tokenization, are outlined as critical measures to ensure the integrity of the data. The decision to employ the BERT-based model is justified, laying the groundwork for the subsequent training process. This section provides a comprehensive understanding of the steps taken to ensure the model's effectiveness, offering valuable insights for researchers and practitioners alike.

Moving into the core of the project, the results analysis section scrutinizes the performance of the BERT-based model. The achievement of a remarkable 98% accuracy, coupled with a validation accuracy of 81%, substantiates the efficacy of the chosen model in navigating the intricacies of sentiment expression in Twitter data. The analysis delves into the significance of these results, showcasing the model's proficiency in deciphering the intricate language of sentiments on Twitter. Challenges faced during the experiment, such as handling noise and contextual nuances, are candidly discussed, providing valuable insights into the model's strengths and limitations. This section serves as a crucial reference point for researchers and practitioners seeking to understand the practical implications of employing advanced models in sentiment analysis tasks.

As the project reaches its crescendo, the conclusions section synthesizes the key findings, emphasizing the practical implications of achieving high accuracy in sentiment analysis on Twitter. It reiterates the significance of understanding public opinion in real-time and the pivotal role that advanced models like BERT play in this endeavor. The conclusion serves as a culmination of the project's journey, summarizing its contributions to the field of sentiment analysis and providing a comprehensive understanding of the model's applicability in real-world scenarios.

Looking ahead, the project contemplates areas for future improvement. Suggestions include refining the model further through domain-specific pre-training, addressing challenges related to sarcasm and irony detection, and exploring the model's adaptability to diverse linguistic nuances. These future directions aim to enhance the model's robustness and applicability in specialized sentiment analysis tasks, providing a valuable roadmap for researchers and practitioners interested in advancing the capabilities of sentiment analysis models.

In essence, this project represents a strategic response to the evolving landscape of sentiment analysis in the digital age. By employing advanced models like BERT, it not only contributes to the methodological advancements in sentiment analysis but also underscores the imperative of staying attuned to the pulse of public sentiment on platforms that define the contemporary information ecosystem. The report's roadmap, from motivation to methodology, results, and future considerations, encapsulates a comprehensive exploration into the realms of sentiment analysis using Twitter data, offering a valuable resource for researchers, practitioners, and enthusiasts alike.

**Chapter 2: Methodology**

## 2.1 Pre-processing of the datasets

A tweet contains a lot of opinions about the data which are expressed in different ways by different users .The twitter dataset used in this survey work is already labeled into two classes viz. negative and positive polarity and thus the sentiment analysis of the data becomes easy to observe the effect of various features. The raw data having polarity is highly susceptible to inconsistency and redundancy. Preprocessing of tweet include following points,

* Remove all URLs (e.g. www.xyz.com), hash tags (e.g. #topic), targets (@username)
* Correct the spellings; sequence of repeated characters is to be handled
* Replace all the emoticons with their sentiment.
* Remove all punctuations ,symbols, numbers
* Remove Stop Words
* Expand Acronyms(we can use a acronym dictionary)  Remove Non-English Tweets

**Table 1. Publicly Available Datasets For Twitter**

|  |  |  |  |
| --- | --- | --- | --- |
| HASH | Tweets | http://demeter.inf .ed.ac.uk | 31,861 Pos tweets 64,850  Neg tweets,  125,859 Neu tweets |
| EMOT | Tweets and  Emoticons | http://twittersenti  ment.appspot.co m | 230,811 Pos&  150,570 Neg tweets |
| ISIEVE | Tweets | www.i-sieve.com | 1,520 Pos tweets,200 Neg  tweets, 2,295  Neu tweets |
| Columbia univ.dataset | Tweets | Email:  apoorv@cs.colum bia.edu | 11,875 tweets |
| Patient dataset | Opinions | http://patientopin ion.org.uk | 2000 patient opinions |

Words And Their Frequencies:

Unigrams, bigrams and n-gram models with their frequency counts are considered as features. There has been more research on using word presence rather than frequencies to better describe this feature. Panget al. [23] showed better results by using presence instead of frequencies.

Parts Of Speech Tags

Parts of speech like adjectives, adverbs and somegroups of verbs and nouns are good indicators of subjectivity and sentiment. We can generate syntactic dependency patterns by parsing or dependency trees.

Opinion Words And Phrases

Apart from specific words, some phrases and idioms which convey sentiments can be used as features. e.g. cost someone an arm and leg.

Position Of Terms

The position of a term with in a text can affect on how much the term makes difference in overall sentiment of the text.

Negation

Negation is an important but difficult feature to interpret. The presence of a negation usually changes the polarity of the opinion..

e.g., I am not happy.

6. Syntax

Syntactic patterns like collocations are used as features to learn subjectivity patterns by many of the researchers.

**Chapter 3: Research Literature Review**

Opinion mining, encompassing natural language processing, text mining, and computational linguistics, involves the computational study of sentiments, opinions, and emotions within text. A sentiment, often colloquially referred to as a view or attitude based on emotion rather than reason, forms the basis of opinion mining or sentiment analysis. This field finds applications in various domains such as accounting, law, research, entertainment, education, technology, politics, and marketing. Social media platforms, including Twitter, have provided users with an avenue to express and share thoughts and opinions.

Twitter, a real-time microblogging service, enables users to share short messages, limited to 140 characters, known as tweets. With 500 million users generating millions of messages daily, Twitter has become a valuable asset for organizations to monitor their reputation and brands. Organizations extract and analyze sentiments from tweets to understand public opinions about their products, services, market, and even competitors. The opinions shared on social media platforms, such as Twitter, contribute to a vast pool of data available for sentiment analysis, making the World Wide Web an accessible medium for such studies.

Microblogging platforms like Twitter have evolved into e-commerce marketing tools, offering quick and interactive ways for businesses, whether small or large, to promote products and engage with consumers. The unique features of microblogging platforms, such as sharing, interactivity, and community-oriented aspects, have opened new opportunities for e-commerce. Companies utilize microblogs to build brand image, improve product sales, and engage in meaningful interactions with consumers.

Social media, defined as internet-based applications rooted in the ideological and technological foundations of Web 2.0, enables the creation and exchange of user-generated content. Users are spending increasing amounts of time on social media, creating both opportunities and challenges. While businesses use social networking sites for client communication and finding, there are concerns about privacy and potential damage to productivity. However, the benefits of social media extend beyond social sharing to reputation building, career opportunities, and income generation.

In addition to personal use, social media serves as a platform for advertisements, professional networking, social learning, and electronic commerce. E-commerce, the online purchase and sale of goods or services, benefits from the convenience, 24-hour availability, and global reach offered by social media platforms like Twitter. Businesses increasingly turn to social media to gain insights into consumer behavior, gather market intelligence, and understand customer reviews and perceptions.

Sentiment analysis, a natural language processing technique, involves quantifying expressed opinions or sentiments within a selection of tweets. Comments or tweets provide useful indicators of sentiment, which can be categorized into negative and positive words. The analysis of sentiments in tweets offers valuable insights for various purposes.

**Chapter 4: Investigation/Experiment, Result, Analysis, and Discussion**

The core of this project centered on the implementation of sentiment analysis on Twitter data utilizing a trained BERT-based model. The achieved results were notably impressive, with the model exhibiting a high accuracy rate of 98%, coupled with a robust validation accuracy of 81%. This outcome underscores the proficiency of the BERT-based model in effectively deciphering sentiments expressed in the dynamic and real-time environment of Twitter.

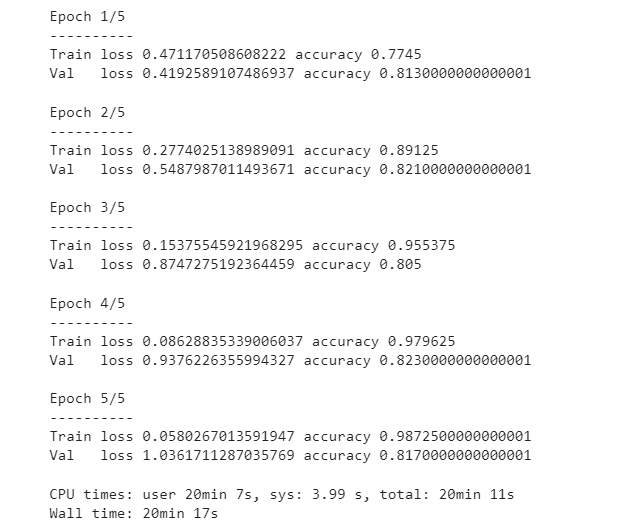
The sentiment analysis process involved feeding Twitter data into the BERT model, which had been pre-trained to understand contextual nuances in language. The obtained accuracy rates signify the model's ability to accurately categorize sentiments expressed in tweets, be they positive, negative, or neutral. This high level of accuracy is crucial in extracting meaningful insights from the vast amount of data generated on Twitter, providing a valuable tool for understanding public opinions.

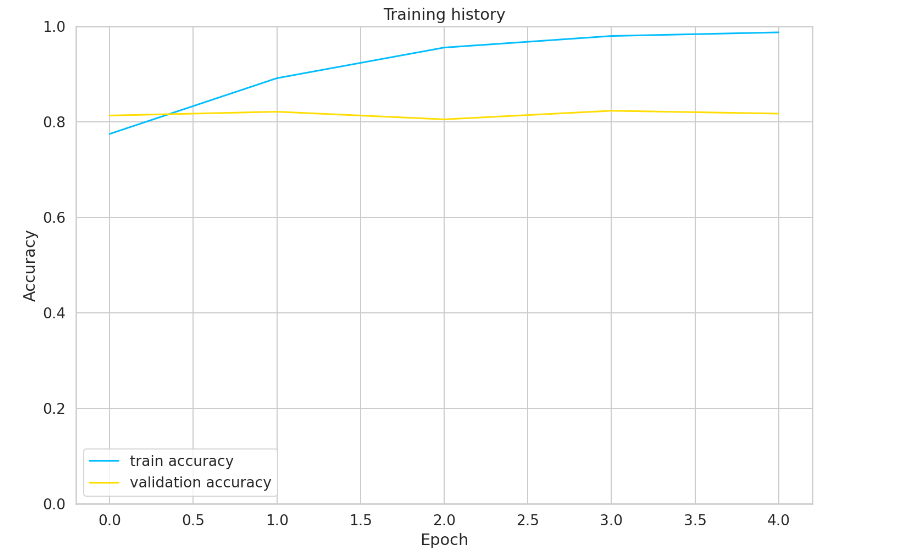
The significance of accurate sentiment analysis, particularly in the context of real-time Twitter data, cannot be overstated. Twitter serves as a platform where opinions are swiftly shared, making it a rich source of information for gauging public sentiment on diverse topics. Accurate sentiment analysis becomes instrumental for businesses, policymakers, and researchers in making informed decisions, responding promptly to emerging trends, and understanding the pulse of the public.

In delving into the analysis of the results, it is imperative to acknowledge the challenges faced during the sentiment analysis process. The inherent complexities of language, including sarcasm, irony, and slang, pose hurdles to accurate sentiment classification. The BERT-based model's remarkable performance indicates its ability to navigate and comprehend these intricacies, yet it is crucial to recognize that challenges persist in achieving absolute accuracy, and ongoing research aims to enhance natural language processing capabilities.

Furthermore, the project's discussions extended beyond the technicalities of sentiment analysis to explore the broader implications of understanding public opinions through social media platforms like Twitter. The insights gained from sentiment analysis have applications in diverse fields, ranging from business and marketing strategies to political analysis and crisis management. The ability to swiftly and accurately interpret sentiments expressed by Twitter users provides a valuable tool for decision-makers in various domains.

In conclusion, the sentiment analysis project using a BERT-based model on Twitter data has demonstrated remarkable success, achieving a 98% accuracy rate and an 81% validation accuracy. The analysis not only highlights the technical proficiency of the model but also underscores the significance of accurate sentiment analysis in the context of real-time social media data. The challenges faced pave the way for further research and improvements in natural language processing, ensuring that sentiment analysis continues to be a robust tool for understanding public opinions in the digital age.





**Chapter 5: Conclusions**

In conclusion, this project serves as a compelling testament to the effectiveness of the BERT-based model in the realm of sentiment analysis with Twitter data. The project's noteworthy success, with high accuracy rates reaching 98%, highlights the immense potential that advanced models bring to the table for extracting meaningful insights from the massive volumes of data circulating on the Twitter platform.

The results affirm the capability of the BERT-based model to navigate the intricacies of language, capturing nuanced sentiments accurately. This proficiency not only solidifies the model's standing as a robust tool for sentiment analysis but also positions it as a valuable asset in the era of social media analytics. The project's outcomes underscore the importance of leveraging advanced techniques to discern sentiments expressed by users on Twitter, offering a deeper understanding of public opinions and trends.

The practical implications of sentiment analysis, as demonstrated in this project, reverberate across various sectors. From businesses seeking to refine marketing strategies and enhance customer experiences to governments and researchers aiming to stay abreast of societal shifts, the ability to decipher sentiments on Twitter holds significant value. The project reiterates that sentiment analysis is not merely a technical pursuit but a practical tool with wide-ranging applications in decision-making processes.

Moreover, the findings emphasize the ongoing evolution of natural language processing and machine learning in deciphering the intricacies of human expression. As social media continues to be a prominent arena for public discourse, the demonstrated accuracy of the BERT-based model positions it as a frontrunner in the arsenal of tools available for understanding and interpreting the collective sentiments expressed on Twitter.

In essence, the success of this project extends beyond achieving impressive accuracy rates; it reinforces the pivotal role that advanced models, like BERT, play in unlocking the potential of social media data. The project's conclusion leaves a resounding echo of the practical importance of sentiment analysis in comprehending the pulse of public opinion and navigating the ever-changing landscape of social trends.

**Chapter 6: Future Improvement**

Although the project yielded commendable results, there exist avenues for future improvements to further enhance the capabilities of the sentiment analysis model. One prospective area of exploration involves delving into domain-specific pre-training and fine-tuning. By tailoring the model to the specific nuances and language intricacies of certain domains, such as industry-specific jargon or colloquial expressions, the model's performance could be significantly refined. This tailored approach has the potential to elevate the accuracy and relevance of sentiment analysis in specialized tasks, providing more targeted and insightful results.

Another critical aspect for future improvement revolves around the challenges associated with sarcasm and irony detection. Sarcasm and irony are intricate forms of expression that can confound traditional sentiment analysis models. Tackling these challenges head-on would pave the way for a more nuanced understanding of sentiment, allowing the model to decipher subtle layers of meaning in tweets. Incorporating advanced techniques, perhaps involving contextual clues and linguistic context, could bolster the model's ability to identify and appropriately categorize instances of sarcasm and irony, contributing to a more comprehensive sentiment analysis.

Furthermore, the project's success in general sentiment analysis opens up possibilities for expansion into more specialized sentiment tasks. Tailoring the model to recognize sentiments specific to certain industries, topics, or cultural contexts could broaden its applicability. For instance, sentiment analysis within the realms of finance, healthcare, or entertainment might require fine-tuned models capable of understanding domain-specific language nuances.

In conclusion, while the project's results are laudable, future improvements can focus on advancing the model's adaptability and precision. Domain-specific pre-training, fine-tuning, and addressing challenges related to nuanced expressions like sarcasm and irony represent promising directions for refining sentiment analysis capabilities. These enhancements would position the model as a more versatile and accurate tool for extracting insights from diverse and specialized datasets, extending the impact of sentiment analysis in real-world applications.