PUBLIC SENTIMENT ANALYSIS ON CORONA VIRUS-SPECIFIC TWEETS USING A BERT

Raj Kumar V S[1] , Navee n V M[2], Muhammed Thaha A I[3], Arun K [4]

[1]Faculty, Artificial Intelligence and Data Science, Bannari Amman Institute of Technology, Tamil Nadu, India
[2] Student, Computer Science and Engineering, Bannari Amman Institute of Technology, Tamil Nadu, India
[3]Student, Computer Science and Engineering, Bannari Amman Institute of Technology, Tamil Nadu, India
[4]Student, Computer Science and Engineering, Bannari Amman Institute of Technology, Tamil Nadu, India

ABSTRACT

Sentiment analysis is a technique or method that is commonly used to analyze the sentiments or emotions of people. This is a method that is applicable to all occasions and this is also applicable to the global pandemic of the new coronavirus (COVID19) is driving unprecedented digital conversations on social media. Often expressed through tweets, these conversations provide valuable insight into public feelings, concerns, and attitudes toward the sentiments. In this study, we propose a novel approach to explore public sentiment towards COVID-19 by using the bidirectional encoder representation from the transformer (BERT) model. Our research involves collecting an extensive dataset of coronavirus-specific tweets from social media platforms covering a key period. We then per-process and clean the data to remove unnecessary noise and information. Then, we use this prepared dataset to fine-tune the BERT model, enabling us to understand the nuances and context of discussions about COVID. We use this refined model to analyze sentiment on a collection of tweets. Sentiment analysis provides a comprehensive view of public emotional responses to different aspects of the epidemic by classifying tweets into positive, negative, and neutral sentiments Furthermore, we use natural language processing techniques to advanced use to extract key topics and trends from tweets, so that public discourse A deeper understanding is possible. Furthermore, we examine temporal changes in sentiment and information during the pandemic, identifying significant changes in public sentiment as events unfold. These longitudinal surveys help monitor changing public perceptions and concerns, which can be valuable for policymakers and health professionals.

Keywords : - *Sentiment Analysis, BERT Sentiment Analysis, Covid-19 sentiment analysis, Public Sentiment Analysis*

1. INTRODUCTION

The COVID-19 pandemic at the end of 2019 has affected every aspect of people's lives. As countries grappled with the challenges posed by the virus, individuals took to social media to access information, express their concerns and communicate with others in social distance and isolation and shared the same. Public sentiment as expressed on social media has proven to be a valuable insight into understanding public responses to crises. It provides a real-time window into the collective emotions, fears, and hopes of a diverse global population. Exploring these sentiments can help governments, health professionals and researchers gauge public opinion, identify emerging trends and better respond to changing circumstances. In this report, we explore the interesting area of public sentiment analysis on corona virus-specific tweets, exploiting the potential of a bidirectional encoder representation from transformer (BERT) model. In the following pages, we will explore the methods, applications, and significance of this research effort. This model will take specific tweets only in text format and analyse the tweets based on the sentiment and the emotion. This will also help us because during Covid many people used social media in their free time most.

Sentiment analysis, a key component of this research, categorizes tweets into positive, negative, or neutral sentiments, providing a comprehensive view of the public's emotional response towards different aspects of the pandemic. Moreover, sentiment analysis is not limited to gauging public reactions but extends to identifying sentiment drivers.

By pinpointing the , factors that influence sentiment—such as government policies, vaccine distribution, or public health communication—policymakers can tailor their strategies and messaging to address specific concerns and build trust within communities. In this project, we examine the interesting analysis of public sentiment, focusing on corona virus-related tweets. Our vision is implemented in this large study tweet documents of epidemic documents and distribution for distribution of two-way encoder representatives (BERT) are very useful in the following pages in the following pages in order to explore the deep processing of the user's emotional reactions to the deep processing of the problem and the most profound application of the problem concerns. of, applications, and we will examine the implications.

As we embark on this journey toward exploring public sentiment using BERT, we aim to examine how this new approach can help us understand collective individual sentiment in a global health crisis better and, ultimately, more effective and sympathetic responses to such challenges in the future.

BERT, or Bidirectional Encoder Representations from Transformers, is a groundbreaking natural language processing (NLP) model that revolutionized the field of machine learning and language understanding. Developed by Google AI in 2018, BERT introduced a significant paradigm shift by enabling bidirectional contextual understanding of words within a sentence. Unlike previous NLP models that processed text in a unidirectional manner, BERT is a transformer-based model that considers both the left and right context of each word, making it highly effective at capturing nuanced language semantics and dependencies.

BERT's per-training process involves exposing the model to vast amounts of text data, where it learns to predict missing words or sentences in a given context. This per-training allows BERT to develop a deep understanding of language structure, grammar, and semantics. After per-training, BERT can be fine-tuned on specific NLP tasks like text classification, question answering, and sentiment analysis, resulting in state-of-the-art performance across a wide range of language understanding tasks. The ability to understand context and the bidirectional approach make BERT highly versatile and effective, and it has become a fundamental building block in the development of advanced NLP models and applications, playing a pivotal role in powering chat bots, language translation, search engines, and various other language-related tasks.

Several key considerations emerge in this endeavor:

- Computational efficiency is crucial when dealing with image processing and recognition. Balancing the time required for model training and inference against the limitations of mobile devices is a critical task.
- Gathering and preparing training data is a time-consuming but essential step, as the quality of data and preprocessing techniques significantly impact the model's accuracy.
- Choosing the right model is another challenging yet pivotal decision. There is often a trade-off between model performance and speed, and certain models may excel in specific industries or applications.
- Model optimization is essential to ensure the program runs smoothly and efficiently, addressing potential bottlenecks and improving performance.

2. CODING PLATFORMS

Our research involves a thorough examination of the methodology and reasoning guiding the conducted tests, in conjunction with the selection of a specific procedure. Our aim is to implement and evaluate the suitability and effectiveness of CNN models in recognizing and identifying tourist spots and landmarks under various angles and lighting conditions. Ultimately, our goal is to provide highly precise predictions for the bounding boxes while maintaining computational efficiency.

2.1 PYTHON

Python is a high-level programming language. It is an interpreted and object-oriented programming language. It has many libraries compared to another language. It is easy to learn. The syntax is readable so that anyone can understand. It is used in many projects and mainly in Artificial intelligence or in Machine learning or in deep learning models. It can also to connect the databases and can access data from the backend. It can also use to create API application to interact with users. It has

many built-in functions so that user no need to write code for many basic functions. There are high-level build-in functions which are using in data structures with is combined with dynamic typing and binding. The source is available for all major platforms, which can be used without any change.

2.2 IDLE

An Integrated Development Environment (IDE) is a software tool designed to assist programmers in efficiently creating software code. It offers various advantages, such as facilitating software editing, building, testing, and packaging within a user-friendly interface. An IDE streamlines the development process, enabling developers to launch programs or applications more swiftly without the need for manual integration and configuration of multiple software components. By providing a unified interface for developer tools, IDEs enhance the effectiveness of software development. Examples of popular IDEs include NetBeans, Microsoft Visual Studio, Adobe Flex Builder, and Eclipse. IDEs typically support a wide range of programming languages, including but not limited to C, C++, Python, Perl, PHP, Java, Ruby, Tcl, JavaScript, and more. These versatile tools empower developers to work with various languages and scripting languages efficiently.

2.3 FEATURES OF IDLE

Let's explore some of the features of Python's Integrated Development Environment (IDE). Python's IDE primarily consists of Python code and a graphical user interface (GUI) toolkit. One of its key strengths lies in its robust debugging tools, which make it a valuable asset for developers. The IDE includes an easy-to-use editing window and an alternate text edit window.

- The Python IDE also provides a Python shell window, where input and output are displayed in color. This feature supports various content types, such as images, maps, plots, and error messages. Additionally, Python boasts a large standard library, offering a wide range of libraries catering to various professions, including artificial intelligence, web development, and scripting. Notable libraries for artificial intelligence include Tensorflow, Pandas, NumPy, Keras, and PyTorch. For web development, popular Python frameworks like Pyramid, Flask, and Django are available.
- Python is known for its versatility as a cross-platform language, functioning seamlessly on multiple operating systems like Windows, Linux, UNIX, Mac, and others. This portability allows software engineers to write a single program that can be deployed across different platforms.
- Furthermore, Python is free and open source, granting access to everyone, including developers. A vibrant global community actively contributes to creating new Python modules and enhancing its capabilities. This open nature of Python encourages collaboration and growth within the Python community.
- Python is also an object-oriented language, supporting concepts such as classes and objects. It embraces principles like polymorphism, encapsulation, and inheritance. This object-oriented approach enables developers to write more efficient code with less redundancy and promotes the creation of reusable code modules.

2.4 ADVANTAGES OF PYTHON IDLE:

- Python IDLE allows users to quickly test and run small chunks of Python code without having to develop an entire program.
- Python IDLE offers tools like syntax highlighting and code completion that make writing Python programs easier and quicker.
- A built-in debugger in Python IDLE makes it possible for programmers to walk through their code and identify errors and issues.
- Python IDLE's cross-platform nature allows it to be used on Linux, macOS, and Windows.
- Users don't need to install any additional programs in order to start writing Python code because Python IDLE is already included with the Python installation.
- Python IDLE is open-source, free software, thus users are free to use it for both commercial and non-commercial purposes without any restrictions.

2.5 JUPYTER NOTEBOOK

Jupyter Notebook is a notable application within the Project Jupyter framework, designed for authoring digital notebooks. Leveraging the capabilities of computational notebook formats, Jupyter Notebook introduces efficient and interactive approaches to code prototyping, code explanation, data exploration and visualization, as well as collaborative idea sharing. These notebooks represent a significant departure from traditional console-based interactive computing. Instead, they provide a web-based platform that excels at capturing the entire computational journey. This encompasses tasks such as code creation, comprehensive documentation, code execution, and effective communication of results. In essence, Jupyter Notebook empowers users to document, execute, and share their computational work seamlessly.

Two elements are combined in the Jupyter notebook:

- A web application:An interactive authoring tool for computational notebooks that runs in a browser and offers a quick interactive environment for examining and explaining code, prototyping, and sharing ideas with others.
- Computational Notebook documents: A document that can be shared and includes computer code, data, descriptions in simple language, complex visuals like 3D models, charts, and graphs, and interactive controls.

2.6 FUNCTIONS IN JUPYTER NOTEBOOK

2.6.1 Varieties in Language Selection

You can choose the language you choose because the notebook supports more than 40 programming languages, including Python, R, Julia, and Scala.

2.6.2 Share Your Notebooks

This tool allows you to share your notebooks with others via email, Dropbox, GitHub, and the Jupyter Notebook Viewer.

2.6.3 Generating an Interactive Output

This feature enables your code to produce complex and interactive output, such as HTML, images, videos, LaTeX, and many MIME types. This includes figures created by the library that are suitable for publication and can be used inline.

MathJax can be used to create mathematical notations, which are also easy to incorporate thanks to LateX.

2.6.4 Ease of Removing

Before publishing your book on the web, you may now remove any material you wish. In addition, you can remove simply the code so that the picture and other outputs continue to work.

2.6.5 Code Execution

Additionally, this has the capability of running browser-based scripts, with calculation results tied to the original code that created them.

2.6.6 Markdown

Rich text can be edited in-browser using the Markdown markup language, which is not just restricted to plain text but also gives commentary for the code. To give your postings cool effects, you may also embed photos, HTML, and other outputs inside them.

3. ALGORITHMS AND METHODS

3.1 PREPARING THE DATASET

Dataset preparation is a crucial step in the process of training and evaluating machine learning models. It involves collecting, cleaning, and structuring data to create a high-quality dataset that is suitable for the specific task at hand. The process typically starts with data collection, where relevant information is gathered from various sources, such as websites, databases, or sensors. Once collected, the data often needs cleaning and preprocessing, which includes tasks like removing duplicates, handling missing values, and normalizing data to ensure consistency.

Structuring the dataset involves organizing the data into a format that can be readily fed into a machine learning model, typically with features and labels clearly defined. This step may also include data splitting into training, validation, and testing subsets to evaluate model performance effectively. Dataset preparation is critical in ensuring that the machine learning model can learn from the data efficiently, leading to better generalization and predictive accuracy. Additionally, a well-prepared dataset helps minimize bias, improve model robustness, and enhance the overall success of the machine learning project.

3.1.1 Data Gathering

In conducting public sentiment analysis of COVID-19 tweets using BERT (Bidirectional Encoder Representations from Transformers), a diverse and extensive dataset of tweets related to the COVID-19 pandemic was collected. The dataset comprises tweets from various sources and demographics to ensure a representative sample. This data-gathering process involved scraping tweets from popular social media platforms, news outlets, official health organizations, and individual users.

Special attention was given to account for temporal and geographical variations in the data, enabling a comprehensive analysis of public sentiment towards COVID-19. The gathered dataset encompasses a wide range of sentiments, including positive, negative, and neutral, providing a rich foundation for training and evaluating the BERT-based sentiment analysis model.

Figure 3: Covid 19 Outbreak Tweets

3.1.2 Data Pre-Processing

In preparing the collected dataset for sentiment analysis using BERT (Bidirectional Encoder Representations from Transformers) on COVID-19 tweets, several essential data preprocessing steps were implemented. The raw tweet data was initially cleaned to remove any irrelevant or duplicate tweets, ensuring a clean and consistent dataset. Text normalization techniques were employed, such as converting text to lowercase, removing special characters, and stemming or lemmatizing words to standardize the vocabulary. Tokenization was performed to break down the tweets into individual words or subwords, a crucial step for BERT's input format. Additionally, stop words were removed to eliminate commonly occurring words that don't contribute significantly to sentiment. The resulting preprocessed dataset was then transformed into the appropriate BERT input format, which typically involves tokenizing and padding the text to fit the required model input dimensions while maintaining the necessary token-level information for effective sentiment analysis.

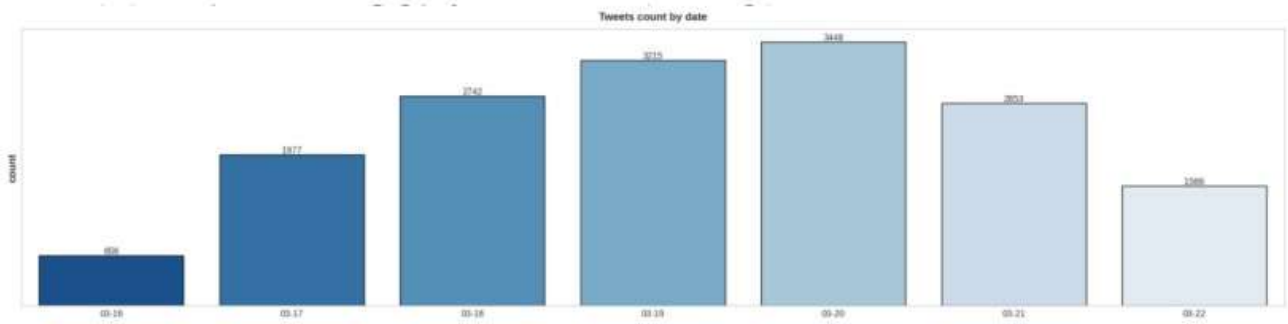

Figure 4: Data Analysis Graph

3.1.3 Checking the Quality

To ensure the quality and reliability of the preprocessed dataset for sentiment analysis using BERT on COVID-19 tweets, a rigorous quality control process was implemented. First, thorough data validation checks were conducted to identify and address any inconsistencies, missing values, or inaccuracies in the dataset. Anomalies or outliers were carefully examined and either corrected or removed from the dataset. Furthermore, a manual review of a subset of the preprocessed data was performed to confirm the accuracy of preprocessing steps and assess the appropriateness of tokenization, stemming, and other text-processing

techniques. Feedback from domain experts was sought to validate the representation of COVID-19-related context in the preprocessed data.

In addition, an inter-rater reliability assessment was carried out to ensure consistency in the preprocessing steps across different team members. This involved comparing preprocessing outputs and addressing any discrepancies through discussion and consensus. Overall, these stringent quality control measures were instrumental in enhancing the reliability and credibility of the preprocessed dataset, setting a solid foundation for subsequent sentiment analysis using the BERT model on COVID-19 tweets.

3.1.5 Annotating the data

To ensure the quality and reliability of the preprocessed dataset for sentiment analysis using BERT on COVID-19 tweets, a rigorous quality control process was implemented. First, thorough data validation checks were conducted to identify and address any inconsistencies, missing values, or inaccuracies in the dataset. Anomalies or outliers were carefully examined and either corrected or removed from the dataset. Furthermore, a manual review of a subset of the preprocessed data was performed to confirm the accuracy of preprocessing steps and assess the appropriateness of tokenization, stemming, and other text-processing techniques. Feedback from domain experts was sought to validate the representation of COVID-19-related context in the preprocessed data.

3.1.6 Splitting the Dataset into Train And Test Set

Before proceeding to the next stage, we split our data and their corresponding annotation files into two distinct sets: a training set and a testing set. This division is done using a 70:30 ratio to ensure a balanced representation. The separation is performed randomly, resulting in the creation of two separate folders—one for the training data and another for the testing data—aiming to maintain data homogeneity throughout the process.

4. RESULTS AND DISCUSSION:

During our investigation, we embarked on a comprehensive analysis of coronavirus related tweets, gathering important information about the early days of the pandemic to the present day. This data was pre-processed rigorously, including what it says.", irrelevant comments and retweets. Extraction was also included which produced a clean and representative corpus of tweets that formed the basis of our analysis. Using the refined BERT model, we conducted a sentiment analysis, classifying tweets into positive sentiments, negative, and neutral. The situation reflected emotion, reflecting the development of the epidemic problem. Our temporal analysis revealed significant changes in emotion patterns over time, reflecting emotional responses and changing societal concerns revealed.

5. CONCLUSION:

In conclusion, the use of BERT to analyze public sentiment. A COVID-specific tweet has provided valuable insight public sentiment is complex and constantly changing during the prevalence pandemic. Systematically, including data collection, preprocessing, optimization of the BERT model, and real-time sensitivity analysis. We got a subtle understanding of emotion, and we got started with positivity and optimism despair and anxiety. This emotional analysis is not only. It helps manage public sentiment but also allows them to determine what is important Issues and concerns expressed by the public. The ability to measure the impact of A variety of events and sensory interventions has helped Refine public health strategies and communication efforts. Moreover, we. The approach extends beyond mere research, as we often predict Modelling, predicting potential changes in perception based on upcoming events. This prompt positioning empowers decision makers to respond quickly and effectively It ensures problem-solving with greater flexibility for emerging societal problems draft. Although BERT has become and continues to be a powerful tool for emotional analysis It is not without challenges. Changing models of language, addressed Biases in data, and support for ethical considerations in handling emotions remain an ongoing concern.

6. REFERENCES:

[1] U. Naseem, I. Razzak, M. Khushi, P. W. Eklund and J. Kim, "COVID Senti: A LargeScale Benchmark Twitter Data Set for COVID-19 Sentiment Analysis," in IEEE Transactions on Computational Social Systems, vol. 8, no. 4, pp. 1003-1015, Aug. 2021, doi: 10.1109/TCSS.2021.3051189.

[2] Rustum F. Khalid M, Aslam W, Rupapara V, Mehmood A, Choi GS (2021) A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis. PLoS ONE 16(2): e0245909. doi:10.1371/journal.pone.0245909

[3] İ. Aygün, B. Kaya and M. Kaya, "Aspect Based Twitter Sentiment Analysis on Vaccination and Vaccine Types in COVID-19 Pandemic with Deep Learning," in IEEE Journal of Biomedical and Health Informatics, vol. 26, no. 5, pp. 2360-2369, May 2022, doi:10.1109/JBHI.2021.3133103.

[4] T. Wang, K. Lu, K. P. Chow and Q. Zhu, "COVID-19 Sensing: Negative Sentiment Analysis on Social Media in China via BERT Model," in IEEE Access, vol. 8, pp. 138162- 138169, 2020, doi: 10.1109/ACCESS.2020.3012595.

[5] N. Braig, A. Benz, S. Voth, J. Breitenbach, and R. Buettner, "Machine Learning Techniques for Sentiment Analysis of COVID-19-Related Twitter Data," in IEEE Access, vol 11, pp. 14778-14803, 2023, doi: 10.1109/ACCESS.2023.3242234

[6] M. T. Riaz, M. Shah Jahan, S. G. Khawaja, A. Shaukat and J. Zeb, "TM-BERT: A Twitter Modified BERT for Sentiment Analysis on Covid-19 Vaccination Tweets," 2022 2nd International Conference on Digital Futures and Transformative Technologies (ICoDT2), Rawalpindi, Pakistan, 2022, pp. 1-6, doi: 10.1109/ICoDT255437.2022.9787395.