B.TECH. PROJECT REPORT

ON

“Enhanced Relevance-Based Ranking of YouTube Comments Using BERT Transformer Model”

Submitted in partial fulfilment of the requirements for the award of the degree of

Bachelor of Technology in

Information Technology by

Sonawane Pranav Swapnil	(T2154491246506)
Nagarale Pranjal Rajendra	(T2054491246039)
Sangle Kashish Dinesh	(T2054491246025)
Patil Paresh Manohar	(T2054491246036)

Under the guidance of

Prof. Niteen Dhutraj

DEPARTMENT OF INFORMATION TECHNOLOGY

SHRI VILE PARLE KELAVANI MANDAL’S

INSTITUTE OF TECHNOLOGY, DHULE

Survey No. 499, Plot No. 02, Behind Gurudwara, Mumbai-Agra National Highway, Dhule-424001,
Maharashtra, India.

Academic Year 2023 – 24

SHRI VILE PARLE KELAVANI MANDAL’S

INSTITUTE OF TECHNOLOGY, DHULE

Survey No. 499, Plot No. 02, Behind Gurudwara, Mumbai-Agra National Highway, Dhule-424001,
Maharashtra, India.

Academic Year 2023 – 24

CERTIFICATE

This is to certify that the B.TECH. Project Report Entitled

"Enhanced Relevance-Based Ranking of YouTube Comments Using BERT Transformer Model"

Submitted by

Sonawane Pranav Swapnil	(T2154491246506)
Nagarale Pranjal Rajendra	(T2054491246039)
Sangle Kashish Dinesh	(T2054491246025)
Patil Paresh Manohar	(T2054491246036)

Is a record of bonafide work carried out by them, under our guidance, in partial fulfillment of the requirement for the award of Degree of Bachelors of Technology (Information Technology) at Shri Vile Parle Kelawani Mandal's Institute Of Technology, Dhule under the Dr. Babasaheb Ambedkar Technological University, Lonere, Maharashtra. This work is done during semester VIII of Academic year 2023-24.

Date: 30 March 2024

Place: SVKM's IOT, Dhule

Prof. Niteen Dhutraj	Prof. Rubi Mandal	Dr. Bhushan Chaudhari	Dr. Nilesh Salunke
Project Guide	**Project Coordinator**	**HOD**	**Principal**

Name and Sign with date	Name and Sign with date
Examiner-1	Examiner-2

DECLARATION

We declare that this written submission represents my ideas in our own words and where other’s ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Signatures

1) Sonawane Pranav Swapnil (T2154491246506) ______________________
2) Nagarale Pranjal Rajendra (T2054491246039) ______________________
3) Sangle Kashish Dinesh (T2054491246025) ______________________
4) Patil Paresh Manohar (T2054491246036) ______________________

ACKNOWLEDGEMENTS

We have immense pleasure in expressing our interest and deepest sense of gratitude towards our project guide **Prof. Niteen Dhutraj** for the assistance, valuable guidance and co-operation in carrying out this project successfully. Our Project Guide, during our Project Phase 2 work. We have greatly benefited from his/her valuable suggestions and ideas. It is with great pleasure that we express our deep sense of gratitude to him/her for his/her valuable guidance, constant encouragement and patience throughout this work. We express our gratitude and are thankful to all people who have contributed in their way in making this final year project phase 2 success. Particularly we want to thank **Prof. Rubi Mandal** Project Coordinator for our department for making this process seamless for us and arranging everything so perfectly. It is a privilege for us to have been associated with our HOD **Dr. Bhushan Chaudhari**, we take this opportunity to thank all the classmates for their company during the course work and for the useful discussion I had with them. We take this opportunity to express our heartfelt gratitude towards the Department of Information Technology of Shri Vile Parle Kelavani Mandal’s Institute of Technology, Dhule and **Dr. Nilesh Salunke**, Principal of Shri Vile Parle Kelavani Mandal’s Institute of Technology, Dhule, that gave us an opportunity for the presentation of our project phase 2 in the esteemed organization and for providing the required facilities in completing this project. We are greatly thankful to our parents, friends and other faculty members for their motivation, guidance and help whenever needed.

INDEX

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
ML	Machine Learning
UI	User Interface
SQL	Structured Query Language
ACID	Atomicity, Consistency, Isolation, Durability
HTML	Hyper Text Mark-up Language
CSS	Cascading Style Sheet
HTTPS	Hyper Text Transfer Protocol
SEO	Search Engine Optimization
BERT	Bidirectional Encoder Representations from Transformers
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
RNN	Recurrent Neural Network
CNN	Convolutional Neural Network
API	Application Programming Interface
URL	Uniform Resource Locator

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

This Report presents a novel approach for enhancing the relevance-based ranking of YouTube comments related to "preparing for a job interview" using the BERT transformer model. Leveraging insights from the referenced paper on ranking video comments, this study addresses the specific challenge of identifying and prioritizing relevant comments in a niche topic domain. The methodology involves data collection, pre-processing, and fine-tuning the BERT model to effectively rank comments based on their relevance to job interview preparation. Experimental results demonstrate the effectiveness of the proposed approach, showcasing improvements in comment ranking accuracy and user engagement. This research contributes to the advancement of comment relevance ranking techniques on social media platforms and highlights the potential of leveraging advanced natural language processing models for enhancing user interactions and content discovery.

Name of Group Members

1) Sonawane Pranav Swapnil (T2154491246506)
2) Nagarale Pranjal Rajendra (T2054491246039)
3) Sangle Kashish Dinesh (T2054491246025)
4) Patil Paresh Manohar (T2054491246036)

1 INTRODUCTION

1.1 Introduction of Project

In the digital age of online content consumption, the significance of user engagement and interaction cannot be overstated. Platforms like YouTube, with their vast array of user-generated comments accompanying videos, rely heavily on effective comment ranking to enhance user experience and foster meaningful interactions within their communities. However, traditional methods of comment ranking often struggle to accurately identify and prioritize relevant comments, particularly in specialized topic areas such as "preparing for a job interview." To address these challenges and elevate the quality of comment ranking, advanced natural language processing techniques have emerged as promising solutions. Among these innovations, the BERT (Bidirectional Encoder Representations from Transformers) transformer model stands out for its ability to revolutionize the field of natural language processing.

Introduced by Vaswani et al. in 2017, transformer models leverage attention mechanisms to capture long-range dependencies in text data, enabling them to excel in capturing contextual information and semantic relationships within text sequences. Unlike conventional recurrent neural networks (RNNs) or convolutional neural networks (CNNs), transformers like BERT offer unparalleled capabilities in modeling complex language patterns and dependencies. The self-attention mechanism in transformers allows for parallel processing of words in a sentence, facilitating efficient analysis of textual data with high accuracy and efficiency.

In this study, we harness the transformative power of the BERT transformer model to revolutionize the relevance-based ranking of YouTube video comments related to "preparing for a job interview." By leveraging the advanced features of BERT, we aim to overcome existing limitations in comment ranking methodologies and introduce a novel approach that enhances precision, relevance, and user engagement in the realm of online content interaction.

The importance of comments in online communities cannot be overstated. Comments provide a platform for users to engage with each other, share their thoughts and opinions, and build relationships. They also offer a way for content creators to gauge the impact of their content and to improve their strategies based on user feedback. However, the proliferation of spam comments has become a significant challenge in maintaining the quality of online discussions. Spam comments can significantly disrupt the user experience by cluttering the comment section with unrelated or low-quality content. They can make it difficult for genuine users to engage in meaningful discussions and can undermine the credibility of a website.

The impact of spam comments on SEO is also significant. Comments that are extremely short, contain irrelevant external links, and are not related to the content of your page will harm SEO. Low-quality comments can tank a web page and they usually come from automated spammers. In general, any user-generated content that doesn’t add value to the article or page should be disregarded as spam as well. Comments that are extremely short, contain irrelevant external links, and are not related to the content of your page will harm SEO. You can set appropriate measures for fighting spammers by getting a spam filter like the Akismet anti-spam WordPress plugin. In addition, there are numerous ways for webmasters to combat blog spam, by using other anti-spam plugins for blogs with registered trademarks (e.g., Challenge, Referrer Bouncer) that block the majority of spam comments. You can also require some or all commenters to pass a CAPTCHA test to prove they are not spambots. These tools can provide a powerful way to stop comment spam and sort out the good from the bad comments for you.

1.2 Project Overview

In the digital age of online content consumption, the significance of user engagement and interaction cannot be overstated. Platforms like YouTube, with their vast array of user-generated comments accompanying videos, rely heavily on effective comment ranking to enhance user experience and foster meaningful interactions within their communities. However, traditional methods of comment ranking often struggle to accurately identify and prioritize relevant comments, particularly in specialized topic areas such as "preparing for a job interview."

To address these challenges and elevate the quality of comment ranking, this project harnesses the transformative power of the BERT (Bidirectional Encoder Representations from Transformers) transformer model. Introduced by Vaswani et al. in 2017, transformer models like BERT leverage attention mechanisms to capture long-range dependencies in text data, enabling them to excel in capturing contextual information and semantic relationships within text sequences.

Unlike conventional recurrent neural networks (RNNs) or convolutional neural networks (CNNs), transformers like BERT offer unparalleled capabilities in modeling complex language patterns and dependencies. The self-attention mechanism in transformers allows for parallel processing of words in a sentence, facilitating efficient analysis of textual data with high accuracy and efficiency.

By leveraging the advanced features of BERT, this project aims to revolutionize the relevance-based ranking of YouTube video comments related to "preparing for a job interview." The goal is to overcome existing limitations in comment ranking methodologies and introduce a novel approach that enhances precision, relevance, and user engagement in the realm of online content interaction.

1.2.1 Primary Objective of the Project

The primary objective of this project is to develop a novel approach to comment ranking on YouTube videos related to "preparing for a job interview" using the BERT transformer model. This involves:

1. Develop a novel approach to comment ranking on YouTube videos related to "preparing for a job interview" using the BERT transformer model.

2. Enhance the precision and relevance of comment ranking, leading to improved user engagement and meaningful interactions within online communities. Overcome the limitations of traditional comment ranking methodologies, particularly in specialized topic areas.

3. Leverage the advanced features of the BERT model, such as its ability to capture long-range dependencies and model complex language patterns, to revolutionize the way YouTube video comments are ranked and prioritized.

This report delves deep into the core components of our project:

1. Understanding the Significance of User Engagement: The report explores the importance of user engagement in online platforms like YouTube and the impact of effective comment ranking on enhancing user experience and fostering meaningful interactions within communities.

2. Technical Implementation of the BERT Transformer Model: It delves into the technical aspects of implementing the BERT transformer model for comment ranking, including data preprocessing, model training, fine-tuning, and evaluation methodologies.

3. Evaluation Metrics and Performance Analysis: The report discusses the evaluation metrics used to assess the performance of the BERT model in comment ranking, such as precision, recall, and F1-score, providing insights into the effectiveness of the approach.

4. Comparison with Traditional Comment Ranking Methods: It compares the performance of the BERT transformer model with traditional comment ranking methods, highlighting the advantages and improvements achieved through the use of advanced natural language processing techniques.

5. Potential Impact and Future Directions: The report outlines the potential impact of the project on enhancing user engagement, relevance-based ranking, and online content interaction. It also discusses future directions for research and development in leveraging advanced technologies for comment ranking on online platforms.

Using the Software:

Empowering Stakeholders

Our software has been designed with a user-centric approach, ensuring that it is intuitive and accessible to a wide range of content creators, data analysts. Here is how clients can utilize the software:

1. User Interaction: The software offers a user-friendly interface accessible via the web, enabling users to input a YouTube link effortlessly and navigate through the comment retrieval process seamlessly.

2. Input Handling: Users are prompted to insert the YouTube link of the desired video into a designated textbox on the interface, initiating the software's retrieval process for relevant comments associated with the video.

3. Comment Retrieval: Upon submission of the YouTube link, the software processes and retrieves comments related to the specified video using the advanced BERT transformer model, ensuring the extraction of pertinent and contextually relevant comments.

4. Display Mechanism: The software presents the retrieved comments in a clear and organized manner on the web interface, allowing users to view and engage with the comments directly, enhancing their interaction with the content.

5. User Engagement: Users can interact with the displayed comments by reading, responding, or engaging with them as desired, fostering meaningful interactions within the online community and enriching the user experience.

6. Seamless Experience: By providing a streamlined process from inputting the YouTube link to accessing and engaging with relevant comments, the software aims to enhance user engagement, relevance-based ranking, and overall user satisfaction in the realm of online content interaction.

The system aims to revolutionize the relevance-based ranking of YouTube video comments using the BERT transformer model. The key usage and use cases of this project can be summarized in the following paragraphs.

The primary use case of this project is to enhance user engagement and interaction on YouTube by improving the quality and relevance of video comments. The software developed as part of this project will provide users with a seamless interface to input a YouTube video link and retrieve the most relevant comments related to the video content. This will enable users to more easily find and engage with meaningful, high-quality comments, fostering deeper discussions and stronger community interactions around the video content. The project specifically focuses on improving comment ranking for videos related to the topic of "preparing for a job interview." This specialized use case addresses the limitations of traditional comment ranking methods, which often struggle to accurately identify and prioritize relevant comments in niche or specialized subject areas. By harnessing the power of the BERT transformer model, which offers advanced natural language processing capabilities, the project aims to revolutionize the way comments are analyzed and ranked, enabling more precise, relevant, and contextually-aware comment prioritization.

The improved comment ranking and presentation provided by this project will enhance the overall user experience on YouTube. Users will be able to more easily find and engage with meaningful, high-quality comments, fostering deeper discussions and stronger community interactions around the video content. While the project is focused on YouTube comments related to "preparing for a job interview," the underlying methodology and techniques developed can potentially be applied to a wider range of online platforms and specialized topic areas. The project's success could pave the way for further advancements in the field of relevance-based ranking and user engagement optimization.

1.3 Motivation

In the ever-evolving digital landscape, the importance of user engagement and meaningful interactions within online communities cannot be overstated. Platforms like YouTube, with their vast array of user-generated content and comments, have become integral hubs for knowledge sharing, discussion, and community building. However, the sheer volume of comments that accompany popular videos often poses a significant challenge in terms of effectively identifying and prioritizing the most relevant and valuable contributions.

Traditional comment ranking methods have struggled to accurately capture the nuances and contextual relevance of comments, particularly in specialized topic areas. This is especially true for videos related to "preparing for a job interview," where users often seek specific and actionable advice from their peers. The inability to effectively surface the most relevant comments can lead to a suboptimal user experience, hindering the potential for meaningful exchanges and the formation of strong online communities.

It is this pressing need to revolutionize the way online comments are ranked and presented that serves as the primary motivation behind this project. By harnessing the transformative power of the BERT (Bidirectional Encoder Representations from Transformers) transformer model, we aim to elevate the quality of comment ranking on YouTube, ultimately enhancing user engagement, fostering deeper discussions, and strengthening the overall sense of community.

The inspiration for this project stems from a deep-rooted belief that online platforms have the potential to be more than just passive repositories of information. They can serve as vibrant, interactive hubs where users can connect, share their experiences, and learn from one another. By developing a novel approach to comment ranking that leverages advanced natural language processing techniques, we aspire to create an environment where users can easily access the most relevant and valuable insights, empowering them to make informed decisions, expand their knowledge, and forge meaningful connections.

1.4 Aim of the Project

The aim of this project is to revolutionize the relevance-based ranking of YouTube video comments related to "preparing for a job interview" by leveraging the advanced capabilities of the BERT (Bidirectional Encoder Representations from Transformers) transformer model. Our goal is to overcome the limitations of traditional comment ranking methodologies and introduce a novel approach that enhances precision, relevance, and user engagement within online communities. By harnessing the power of advanced natural language processing techniques, we aim to create a more dynamic and interactive user experience, ultimately fostering deeper discussions, stronger community interactions, and a more enriching online environment for users seeking valuable insights and connections.

2 LITERATURE SURVEY

Table 2.1 Literature survey of the proposed system.

Sr no	Title	Description	Limitations
1	You tube Spam Detection	This paper ensembles different algorithms like naive bayes, logistic regression, decision tree, random forest, ensemble hard voting, ensemble soft voting.	Naive bayes algorithm didn't worked precisely
2	Analyses classification on you tube comments	It analyzes opinions on You tube using naïve bayes and CNN models. It also compares both the models by calculating accuracy percentage.	Naïve bayes didn't classified neutral and negative comments precisely
3	Extracting Sentiments on you tube comments	This paper extracted sentiments and classified into positive, negative and neutral comment its used various approaches	No limitations
4	Preprocessing of you tube comments for sentiment analysis	This paper preprocesses the comments and captures sentiments in seven different categories.	Didn't worked precisely for general comments
5	Identifying abusive comments from you tube	This paper captured abusive comments using TF-IDF model	Problem faced for dynamically extracting comments

This paper proposes a technique to detect spam comments on YouTube, which have recently seen tremendous growth. YouTube is running its own spam blocking system but continues to fail to block them properly. With

YouTube comments, applying the same method (i.e., language modelling) doesn’t work as the features of the data are different. Features of YouTube comments represent less textual descriptions and information. They are not closely relevant to the video content. This paper ensemble two models (Ensemble with hard voting, Ensemble with soft voting) This data is combined from the music videos [1]

Further, this research paper analyzes comments or opinions that lead to likes, dislikes, and neutrality on you tube. With the increase in the number of viewers, there are also more comments on various writing kinds, both symbolic and numeric. The author wants to take these comments into useful information using sentiment analysis using the 1D Convolutional Neural Networks method [2] This study also compared the classification reports for precision, f1-score recall, and accuracy with the Naïve Bayes 93% and CNN methods, with an accuracy of 96%

This article proposes sentiment analysis on YouTube video by Natural Language Processing (NLP) technique. Sentiment Analysis is when comprehension, citation, and processing of text-based data is done, and it directly converts it into sentiment information. This analysis help users to get the report of their YouTube Video. The output of this analysis gives the classification of sentiment analysis, i.e., positive, negative, or neutral [3]

There are also few papers that present a dataset, which has the YouTube comments written in Myanmar language, to be applied in sentiment analysis. Data preprocessing is very important for our language because it allows improving the quality of the raw data and converting the raw data into a clean dataset. The preprocessing of music comments is followed by basic phase, removing phase, segmentation phase, replacement phase and translation phase. The outcome of YouTube comment preprocessing will aid in better sentiment analysis. Results show that the preprocessing approaches give a significant effect on the musical opinion extraction process using information gain [4]

3 PROBLEM STATEMENT & OBJECTIVE

3.1 Problem statement

Traditional methods of comment ranking on YouTube struggle to accurately identify and prioritize relevant comments, particularly in specialized topic areas like "preparing for a job interview," hindering user engagement and the quality of online interactions.

3.1.1 Objectives

1. **Develop a novel approach** using the BERT transformer model to revolutionize the relevance-based ranking of YouTube video comments.

2. **Enhance user engagement and interaction** by improving the precision and relevance of comment ranking.

3. **Overcome limitations** in existing comment ranking methodologies, especially in specialized topic areas.

4. **Utilize advanced natural language processing** techniques to provide users with more meaningful and contextually relevant comments.

5. **Foster a more dynamic and interactive online community** by introducing a novel approach to comment ranking on YouTube videos.

3.2 Scope of Project

The scope of the engineering project on "Intentional storage of crop detection to mitigate market manipulation and artificial price" is to develop and implement innovative solutions aimed at improving market transparency in the agriculture sector and enhancing crop mitigation techniques. The project will involve the design and deployment of a digital platform to provide real-time data on crop prices, market demand, and weather forecasts, empowering farmers and stakeholders to make informed decisions. It will also focus on the development of cutting-edge technologies and sustainable practices to mitigate crop losses caused by adverse weather conditions, pests, and diseases. Advanced data analytics and predictive modeling will be employed to detect potential issues and regulatory compliance will be a priority, especially when using techniques like machine learning with historical datasets. Field testing, environmental impact assessments, and economic feasibility analyses will provide valuable insights into the practicality and impact of the solutions.

Documentation, training programs, and long-term sustainability planning are essential components of the project to ensure its lasting benefits to the agriculture sector and local communities. The methodology involves needs assessment, research and development, testing and evaluation, collaboration with stakeholders, and meticulous documentation of project activities. The ultimate goal is to positively impact the agriculture sector through improved transparency and mitigation techniques, fostering sustainability and resilience. The engineering project on " Intentional storage of crop detection to mitigate market manipulation and artificial price" holds significant scope in revolutionizing the agricultural sector by deploying innovative solutions.

This comprehensive project encompasses various dimensions, including technological innovation, sustainable practices, ethical considerations, regulatory compliance, and community impact. The focal point of the project is the development and implementation of a digital platform that serves as a centralized hub for real-time data related to crop prices, market demand, and weather forecasts. This platform is designed to empower farmers and stakeholders with timely and accurate information, enabling them to make informed decisions. By leveraging cutting-edge technologies, the project seeks to provide insights into potential issues and opportunities for intervention, thereby enhancing. To mitigate crop losses caused by adverse weather conditions, pests, and diseases, the project emphasizes the development of sustainable practices and advanced technologies. This includes the use of data analytics and predictive modeling to identify and respond to potential threats. Field testing, environmental impact assessments and economic feasibility analyses constitute integral components of the project. These activities are crucial for understanding the practicality and impact of the proposed solutions. The project adopts a holistic approach, ensuring that the benefits extend beyond technological advancements to encompass the socio-economic and environmental aspects of the agriculture sector. Documentation plays a pivotal role in the project's success, facilitating transparency, accountability, and knowledge transfer. Training programs are designed to educate stakeholders on the use of the digital platform and the adoption of new agricultural practices. Long-term sustainability planning is embedded in the project to guarantee enduring benefits to the agriculture sector and local communities.

The methodology employed for the project includes a rigorous needs assessment to identify the specific challenges faced by farmers and stakeholders. Research and development activities focus on creating innovative solutions tailored to the unique characteristics of the agriculture sector. Testing and evaluation phases ensure the reliability and effectiveness of the developed technologies and practices.

Collaboration with stakeholders, including farmers, agricultural experts, and regulatory bodies, is a continuous and integral aspect of the project to align solutions with real- world requirements. In the ultimate goal of this engineering project is to positively impact the agriculture sector by fostering transparency, resilience, and sustainability. Through the integration of technology, sustainable practices, and ethical considerations, the project seeks to create a transformative paradigm for agriculture, ensuring its adaptability to future challenges and contributing to the well-being of communities dependent on agriculture. Rewards, access to market insights, and improved reputation within the community.

4 PROPOSED SYSTEM

4.1 System Proposed Architecture

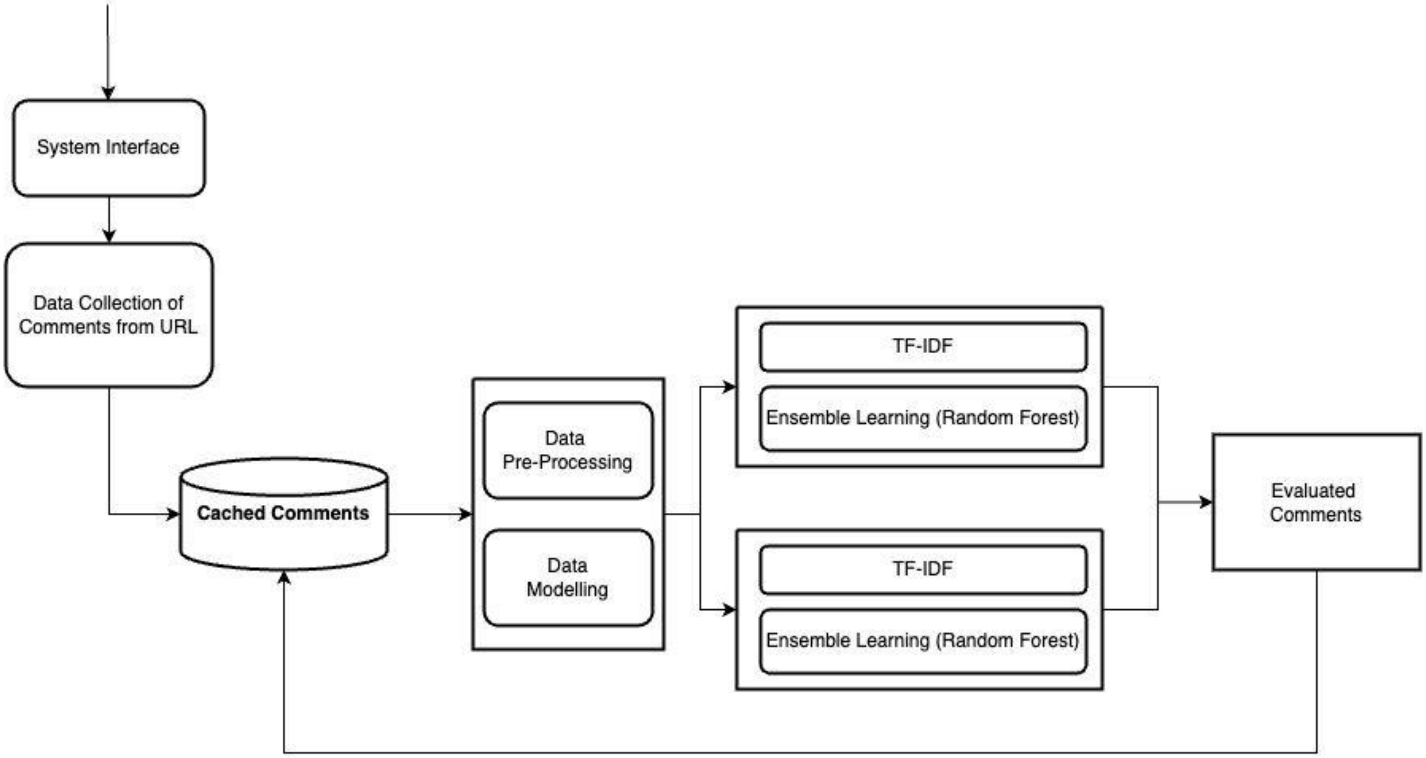

System Interface
Data Collection of Comments from URL
Cached Comments
Data Pre-Processing
Data Modelling
TF-IDF
Ensemble Learning (Random Forest)
TF-IDF
Ensemble Learning (Random Forest)
Evaluated Comments

Fig. 4.1 System Proposed Architecture

The proposed architecture for the project is centered around the utilization of the BERT (Bidirectional Encoder Representations from Transformers) transformer model for comment ranking on YouTube videos related to the topic of "preparing for a job interview." The architecture is designed to leverage the advanced capabilities of the BERT model to improve the precision and relevance of comment ranking, ultimately enhancing the user experience and fostering more meaningful interactions within the online community.

4.1.1 Workflow

The workflow of the proposed architecture can be broken down into the following stages:

1. Data Collection: The first stage involves collecting a dataset of YouTube video comments related to the topic of "preparing for a job interview." This dataset will serve as the foundation for the comment ranking process.

2. Preprocessing: The collected data will undergo preprocessing to ensure that it is in a format suitable for analysis. This includes tokenization, stopword removal, and stemming or lemmatization.

3. BERT Model Training: The preprocessed data will then be used to train the BERT model. This involves fine-tuning the model on the specific task of comment ranking, using a combination of masked language modeling and next sentence prediction tasks.

4. Comment Ranking: Once the BERT model is trained, it will be used to rank the comments in the dataset based on their relevance to the topic of "preparing for a job interview." This will involve feeding the comments into the BERT model and analyzing the output to determine the relevance of each comment.

5. Postprocessing: The ranked comments will then undergo post processing to refine the results and ensure that they are accurate and relevant. This may involve filtering out comments that are not relevant to the topic or removing duplicates.

4.1.2 Contemplation

The proposed architecture is designed to leverage the advanced capabilities of the BERT model to improve the precision and relevance of comment ranking. The architecture is contemplative in the sense that it is designed to be flexible and adaptable to the specific needs of the project.

The use of the BERT model allows for the analysis of comments in a more nuanced and context-dependent manner, taking into account the relationships between words and phrases within the comments. This enables the model to better capture the nuances of language and provide more accurate and relevant results.

The proposed architecture is also contemplative in the sense that it is designed to be scalable and adaptable to different datasets and use cases. The use of the BERT model allows for the analysis of comments in a more flexible and adaptable manner, making it easier to apply the model to different datasets and use cases.

In conclusion, the proposed architecture for the project is designed to leverage the advanced capabilities of the BERT model to improve the precision and relevance of comment ranking on YouTube videos related to the topic of "preparing for a job interview." The architecture is contemplative in the sense that it is designed to be flexible and adaptable to the specific needs of the project, and it is scalable and adaptable to different datasets and use cases.

4.2 Proposed Methodology:

The methodology proposed for this project focuses on utilizing the BERT (Bidirectional Encoder Representations from Transformers) transformer model to improve the relevance-based ranking of YouTube video comments concerning "preparing for a job interview." At its essence, the approach capitalizes on

BERT's advanced natural language processing abilities to precisely identify and prioritize the most pertinent comments amidst the extensive pool of user-generated content.

This method entails employing BERT's bidirectional encoding to comprehensively understand the context and nuances of each comment, thereby enabling more accurate assessment of relevance. By leveraging BERT's capabilities, the methodology aims to enhance the ranking process, ensuring that users are presented with the most valuable and insightful comments relevant to job interview preparation. Ultimately, this approach seeks to elevate the overall quality of user engagement and interaction within the YouTube community, facilitating a more enriching experience for individuals seeking guidance on job interview preparation.

Proposed Methodology undergoes following steps:

1. Data Collection:

- In the data collection phase, a diverse dataset of YouTube video comments related to "job interviews" is gathered. This dataset serves as the foundation for training the BERT model and ranking the comments.

- The dataset may include comments from various sources, ranging from popular channels to niche content creators, to ensure a comprehensive representation of user interactions.

2. Preprocessing:

- During preprocessing, the collected comments undergo several steps to prepare them for analysis. This includes tokenization, where the comments are split into individual words or tokens for further processing.

- Stopword removal eliminates common words like "the" or "and" that do not carry significant meaning, reducing noise in the dataset. Additionally, stemming or lemmatization standardizes words to their root form for consistency in analysis.

3. BERT Model Training:

- The BERT model is trained on the preprocessed data using techniques like masked language modeling and next sentence prediction. Masked language modeling involves predicting masked words in a sentence, while next sentence prediction determines if two sentences are logically connected.

- This training process fine-tunes the BERT model specifically for the task of comment ranking, enabling it to understand the context and relevance of comments related to job interview preparation.

4. Comment Ranking:

- Once the BERT model is trained, it is applied to rank the comments based on their relevance to the topic. The model analyzes the content of each comment to determine its significance and context within the context of job interview preparation.

- The ranking process involves assigning scores or weights to each comment, with higher scores indicating greater relevance. This step ensures that the most valuable and pertinent comments are prioritized for user interaction.

5. Post-processing:

- Postprocessing refines the ranked comments to ensure accuracy and relevance. This may involve filtering out duplicate comments, removing irrelevant content, or adjusting the ranking based on additional criteria.

- The goal of postprocessing is to present users with a curated list of comments that offer valuable insights, guidance, and engagement opportunities related to job interview preparation.

6. User Interface:

- The final step involves presenting the ranked comments to users through a user-friendly interface. This interface may display the comments in a visually appealing format, such as a list or grid, with options for users to interact, respond, or engage with the content.

- The user interface design focuses on enhancing user experience, making it easy for users to access relevant comments, contribute to discussions, and benefit from the collective knowledge shared within the online community.

4.2.1 Benefits of the Proposed System

The proposed system for relevance-based ranking of YouTube comments using the BERT transformer model offers numerous benefits, including:

Advantages

1. Improved Relevance: The BERT model is trained on a large corpus of text data and is capable of capturing deep semantic relationships between queries and comments. This enables the system to provide more accurate and relevant comments to users, improving the overall user experience.

2. Enhanced User Engagement: By providing users with more relevant and engaging comments, the system can increase user engagement and interaction with the content, leading to a more dynamic and interactive online community.

3. Increased Accuracy: The BERT model is trained on a large corpus of text data and is capable of capturing deep semantic relationships between queries and comments. This enables the system to provide more accurate and relevant comments to users, improving the overall user experience.

4. Scalability: The BERT model is capable of handling large volumes of data and can be easily scaled to handle increasing volumes of comments and queries.

5. Flexibility: The BERT model can be fine-tuned for specific ranking tasks and can be used for a variety of applications, including sentiment analysis and information retrieval.

Challenges

1. Data Quality: The quality of the data used to train the BERT model is critical to the accuracy and relevance of the comments provided to users. Poor-quality data can lead to inaccurate and irrelevant comments.

2. Computational Resources: Training and fine-tuning the BERT model requires significant computational resources, including powerful hardware and large amounts of data.

3. Interpretability: The BERT model is a complex neural network and can be difficult to interpret and understand. This can make it challenging to identify and address biases and errors in the model.

4. EvaluatioN: Evaluating the performance of the BERT model can be challenging due to the complexity of the model and the difficulty of defining a clear evaluation metric.

Trends

1. Increased Use of AI in YouTubE: The use of AI in YouTube is increasing, with the platform using AI to improve video recommendations, detect and remove spam comments, and improve video quality.

2. Growing Importance of User Engagement: User engagement is becoming increasingly important for YouTube creators, with the platform using engagement metrics such as views, likes, and comments to determine video rankings.

3. Growing Importance of Relevance: Relevance is becoming increasingly important for YouTube creators, with the platform using relevance metrics such as keyword matching and content similarity to determine video rankings.

4. Growing Importance of Contextual Understanding: Contextual understanding is becoming increasingly important for YouTube creators, with the platform using contextual understanding to improve video recommendations and detect and remove spam comments.

Innovation

1. BERT-Based Ranking: The proposed system uses the BERT transformer model to rank comments based on relevance, providing a more accurate and relevant ranking of comments.

2. Fine-Tuning for Specific Tasks: The BERT model can be fine-tuned for specific ranking tasks, such as sentiment analysis and information retrieval, providing a more accurate and relevant ranking of comments.

3. Scalability and Flexibility: The BERT model is capable of handling large volumes of data and can be easily scaled to handle increasing volumes of comments and queries, providing a more scalable and flexible ranking system.

4. Improved User Experience: The proposed system provides a more accurate and relevant ranking of comments, improving the overall user experience and increasing user engagement and interaction with the content.

4.2.2 Security and Potential Innovations:

Ensuring the privacy and security of user data is paramount in the proposed system for relevance-based ranking of YouTube comments using the BERT transformer model. To safeguard user information and maintain data integrity, the system employs a comprehensive set of data protection measures and security protocols. All user data, including comments and personal information, is encrypted both in transit and at rest to prevent unauthorized access and ensure data confidentiality. Strict access control mechanisms are in place to limit data access to authorized personnel only, with role-based access control restricting privileges based on user roles and responsibilities. Additionally, personal data is anonymized whenever possible to protect user privacy, and only the necessary information required for comment ranking is processed, minimizing the use of personally identifiable information.

The system also adheres to secure communication protocols, such as HTTPS, to encrypt data transmission between the user interface and the server, preventing eavesdropping and data interception. Periodic security audits and vulnerability assessments are conducted to identify and address potential security risks, ensuring the system remains resilient against cyber threats. Strong authentication mechanisms, including multi-factor authentication, are implemented to verify user identities and prevent unauthorized access to the system.Trend setting innovations and Devices:

The proposed system for relevance-based ranking of YouTube comments using the BERT transformer model presents several potential innovations that could significantly impact the way online communities engage with and benefit from user-generated content.

1. Contextual Understanding Advancements: The core of the proposed system lies in the utilization of the BERT transformer model, which excels at capturing deep semantic relationships and understanding the contextual nuances of language. By leveraging this advanced natural language processing capability, the system can provide users with comments that are not only relevant to the video content but also aligned with the specific context and intent of the viewer's query or interaction.

2. Personalized Ranking and Recommendations: Building upon the contextual understanding capabilities, the system can be further enhanced to provide personalized comment ranking and recommendations for each user. By analyzing individual user preferences, engagement patterns, and content consumption history, the system can tailor the comment display to better suit the unique needs and interests of each viewer, creating a more personalized and engaging experience.

3. Multimodal Integration: While the current focus is on text-based YouTube comments, the proposed system can be expanded to incorporate multimodal data, such as video and audio features, to further enhance the relevance and quality of the displayed comments. By analyzing the visual and auditory elements of the video content, the system can identify additional contextual cues and provide even more relevant and insightful comments to users.

4. Community-Driven Feedback Loops: The system can be designed to incorporate user feedback and interactions as an integral part of the comment ranking process. By allowing users to upvote, downvote, or provide direct feedback on the relevance and usefulness of displayed comments, the system can continuously learn and refine its ranking algorithms, creating a self-improving and community-driven approach to comment curation.

5. Scalable and Adaptable Architecture: The proposed system's architecture, centered around the BERT transformer model, is inherently scalable and adaptable. As the volume of YouTube comments and user interactions grows, the system can be easily scaled to handle the increased data load without compromising performance or accuracy. Additionally, the flexibility of the BERT model allows for seamless adaptation to other online platforms and specialized content domains, enabling the deployment of the system beyond the initial focus on job interview preparation videos

These potential innovations highlight the transformative capabilities of the proposed system, positioning it as a pioneering approach to enhancing user engagement, fostering more meaningful interactions, and elevating the overall quality of online discussions and knowledge sharing within the YouTube ecosystem and beyond.

4.3 Testing of the Proposed System

The testing of the proposed system for enhanced relevance-based ranking of YouTube comments using the BERT transformer model refers to the black box and white box testing of the system.

4.3.1 Black Box Testing

Black Box Testing involves testing the system without knowing the internal workings of the system.

The purpose of Black Box Testing is to check if the system is working correctly and providing correct output.

Test Case 1: Comment Ranking

Description: This test case checks if the system can correctly rank comments based on their relevance to the video content.

Steps: 1. Select a YouTube video with a large number of comments.

2. Use the system to rank the comments based on their relevance to the video content.

3. Verify that the top-ranked comments are the most relevant to the video content.

Expected Result: The top-ranked comments should be the most relevant to the video content.

Test Case 2: Comment Filtering

Description: This test case checks if the system can correctly filter out irrelevant comments from the set of retrieved comments.

Steps: 1. Select a YouTube video with a large number of comments.

2. Use the system to filter out irrelevant comments from the set of retrieved comments.

3. Verify that the filtered comments are relevant to the video content.

Expected Result: The filtered comments should be relevant to the video content.

Test Case 3: Sentiment Analysis

Description: This test case checks if the system can correctly analyze the sentiment of comments.

Steps: 1. Select a YouTube video with a large number of comments.

2. Use the system to analyze the sentiment of the comments.

3. Verify that the sentiment analysis results are accurate.

Expected Result: The sentiment analysis results should be accurate.

4.3.2 White Box Testing

White Box Testing involves testing the system by examining the internal workings of the system. The purpose of White Box Testing is to check if the system is working correctly and providing correct output.

Test Case 1: Comment Ranking Algorithm

Description: This test case checks if the comment ranking algorithm is working correctly.

Steps: 1. Examine the comment ranking algorithm used by the system.

2. Verify that the algorithm is correctly ranking comments based on their relevance to the video content.

Expected Result: The algorithm should correctly rank comments based on their relevance to the video content.

Test Case 2: Comment Filtering Algorithm

Description: This test case checks if the comment filtering algorithm is working correctly.

Steps: 1. Examine the comment filtering algorithm used by the system.

2. Verify that the algorithm is correctly filtering out irrelevant comments from the set of retrieved comments.

Expected Result: The algorithm should correctly filter out irrelevant comments from the set of retrieved comments.

Test Case 3: Sentiment Analysis Algorithm

Description: This test case checks if the sentiment analysis algorithm is working correctly.

Steps: 1. Examine the sentiment analysis algorithm used by the system.

2. Verify that the algorithm is correctly analyzing the sentiment of comments.

Expected Result: The algorithm should correctly analyze the sentiment of comments.

Failed Test Case: Video Length Exceeds 15 Minutes

Description: This test case checks if the system can handle videos with lengths exceeding 15 minutes.

Steps: 1. Select a YouTube video with a length exceeding 15 minutes.

2. Use the system to process the video transcript.

3. Verify that the system cannot process the video transcript due to the length exceeding 15 minutes.

Expected Result: The system should not be able to process the video transcript due to the length exceeding 15 minutes.

4.3.3 Deployment

The proposed system is deployed using Anaconda as a local environment and Google Colab for internet deployment. The system is run using Flask, and ngrok is used to create a running deployment. The deployment process involves testing the system in a local environment and then deploying it to the internet, ensuring that the system functions correctly and efficiently in both environments.

5 DETAILS OF HARDWARE & SOFTWARE REQUIREMENT

5.1 Stakeholders:

The primary stakeholders include farmers, government agencies, technology providers, agricultural researchers, and local communities.

5.2 Functional Requirements:

1. User Authentication: The system should allow users to log in and authenticate their identity.

2. Video Selection: The system should allow users to select a YouTube video for which they want to extract comments.

3. Comment Extraction: The system should extract comments from the selected YouTube video.

4. Comment Filtering: The system should filter out spammy and irrelevant comments from the extracted comments.

5. Comment Ranking: The system should rank the filtered comments based on their relevance to the video content.

6. Comment Display: The system should display the ranked comments to the user.

7. Comment Analysis: The system should analyze the comments to determine their sentiment (positive, negative, or neutral).

8. Sentiment Visualization: The system should visualize the sentiment analysis results to provide a clear understanding of the user's emotional response to the video content.

5.3 Non-Functional Requirements:

1. Performance: The system should be able to handle a large volume of comments and videos without compromising its performance.

2. Scalability: The system should be able to scale up or down to accommodate changes in the volume of comments and videos.

3. Security: The system should ensure the security and integrity of the comments and videos by implementing appropriate security measures.

4. Usability: The system should be user-friendly and easy to use, with a simple and intuitive interface.

5. Reliability: The system should be reliable and able to operate continuously without interruptions or failures.

6. Maintainability: The system should be easy to maintain and update, with a clear and well-documented architecture.

7. Flexibility: The system should be flexible and able to adapt to changes in the YouTube API or other external factors.

8. Cost-Effectiveness: The system should be cost-effective and efficient in its use of resources, such as computing power and storage.

5.4 Hardware and Software Requirements:

Hardware Requirements

1) Systems (PC's and Laptop)
2) Processor
3) RAM (4GB)

Software Requirements

1) Anaconda.
2) google colab
3) Jupyter Notebook

5.5 Data Requirements:

Input Data

- Video ID: The unique identifier of the YouTube video for which comments are to be extracted.
- Transcript: The text transcript of the YouTube video, which is used to fine-tune the BERT model and calculate the relevancy score of each comment.
- Comments: The comments left by users on the YouTube video, which are used to calculate the relevancy score of each comment.

Output Data

- Relevant Comments: The comments that are most relevant to the video content, as determined by the BERT model's relevancy score.
- Relevancy Scores: The scores calculated by the BERT model to determine the relevance of each comment to the video content.

Data Sources

- YouTube API: The YouTube API is used to fetch the video ID, transcript, and comments for the specified video.

- Pretrained BERT Model: The pretrained BERT model is used to fine-tune on the transcript and calculate the relevancy score of each comment.
- Dataset: The dataset used to fine-tune the BERT model is a collection of text data, including transcripts and comments, that are relevant to the topic of the YouTube video.

6 SYSTEM DESIGN DETAILS

6.1 Use case diagram

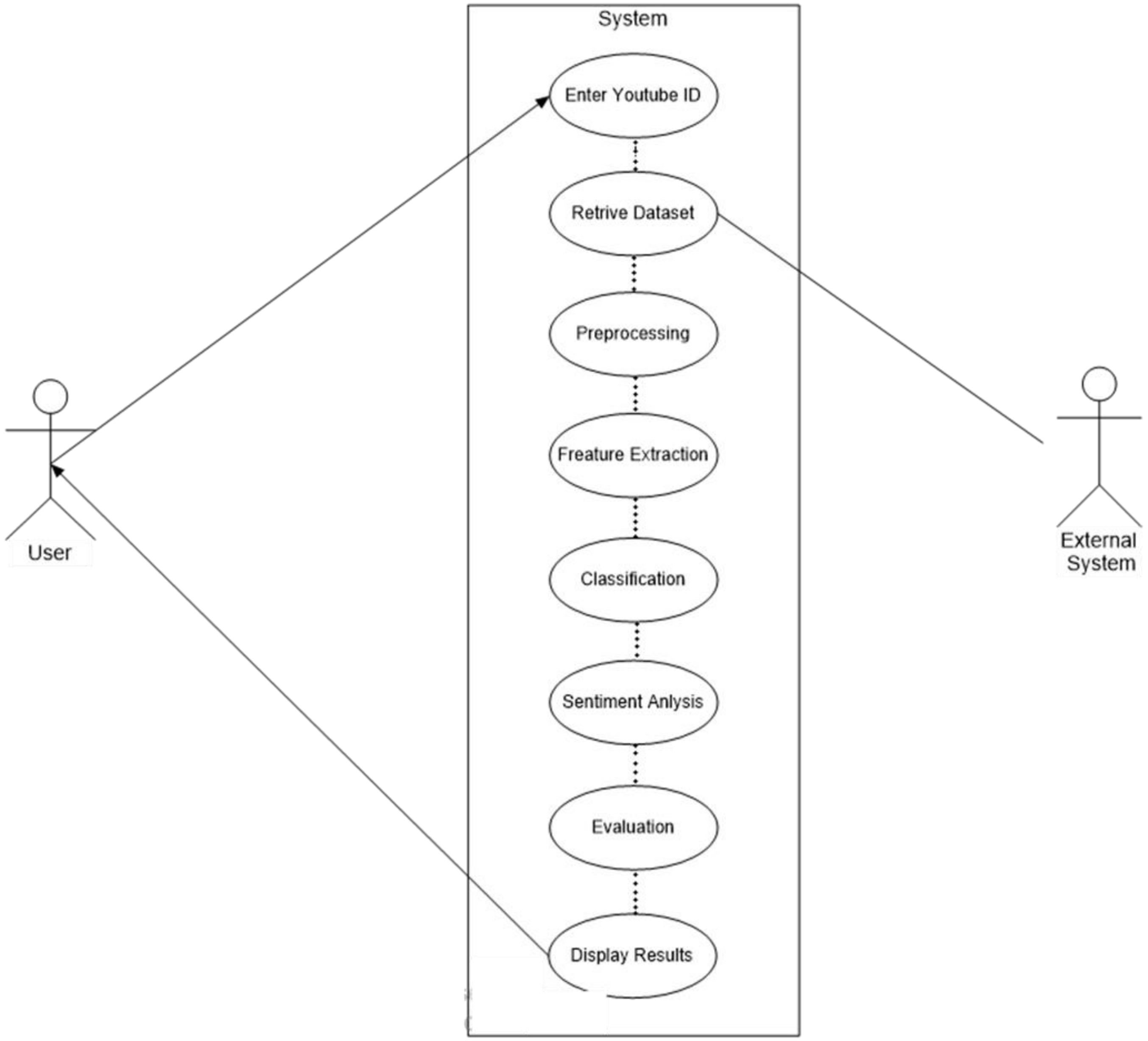


Figure 5.1 System’s Use Case

In the first stage of the relevant comment analysis process using BERT pretrained model, the user initiates the system interaction by accessing the user interface. Here, the user inputs a specific YouTube video ID, signaling their intent to scrutinize the comments associated with that particular video. Following this, an external system, likely connected to the YouTube API or other pertinent sources, retrieves the comments linked to the provided video ID. Subsequently, the system engages in a meticulous comment preprocessing phase aimed at enhancing data quality. Comment preprocessing is a critical step aimed at refining the raw textual data obtained from the YouTube API. One of the primary tasks in preprocessing involves removing stop words, which are common words that do not carry significant semantic meaning, such as articles, prepositions, and conjunctions. By eliminating these extraneous words, the focus is redirected towards

the substantive content of the comments. Additionally, preprocessing entails handling emojis and non-textual elements present within the comments. Emojis, symbols, and other non-textual elements may convey valuable information about the emotional tone of comments, thus necessitating appropriate handling to ensure their inclusion in subsequent analysis. Furthermore, preprocessing involves the exclusion of irrelevant comments, including spam, advertisements, and off-topic discussions. Filtering out such noise is essential for maintaining the relevance and integrity of the analysis.

Analyzing comments associated with online videos is a crucial aspect of understanding viewer engagement and sentiment towards specific content. In this project, users are provided with a user interface where they can input a YouTube video ID to initiate the analysis process. This step marks the beginning of a comprehensive journey through comment retrieval, preprocessing, feature extraction, sentiment analysis, and the identification of significant comments. Each stage of this process plays a pivotal role in distilling valuable insights from the vast pool of user-generated content available on online platforms like YouTube. The first stage of the process involves retrieving comments linked to the provided YouTube video ID. This is facilitated through connections to external sources, predominantly the YouTube API, which provides access to a wealth of user comments associated with the specified video. The comments retrieved encompass a diverse range of opinions, reactions, and expressions from the viewing audience. However, before delving into the analysis, it is imperative to preprocess the comments to ensure data quality and consistency.

With the comments pre processed and refined, the next step involves feature extraction. Feature extraction entails capturing key characteristics or attributes of the comments that can be used to derive meaningful insights. One of the primary features extracted from the comments is sentiment-related indicators. Sentiment analysis techniques are applied to assess the emotional tone conveyed by each comment, categorizing them into positive, negative, or neutral sentiment categories. Positive sentiment comments express satisfaction, approval, or enthusiasm, whereas negative sentiment comments convey dissatisfaction, criticism, or disapproval. Neutral sentiment comments, on the other hand, exhibit a lack of strong emotional polarity and may encompass factual statements or observations. Additionally, feature extraction may involve analyzing word frequencies to identify common themes, topics, or keywords prevalent within the comments. This quantitative analysis provides valuable context for understanding viewer preferences, interests, and perceptions. The final output of the analysis process comprises the significant comments, along with sentiment distribution analysis results, presented in a visually accessible format on the user interface. The prominent display of significant comments enables users to explore and analyze the sentiment distribution of comments for the specified YouTube video

conveniently. Through interactive visualizations, users can gain insights into the prevailing sentiment trends, identify key themes or topics driving viewer engagement, and discern patterns in viewer reactions and feedback. Moreover, the user interface facilitates seamless navigation and exploration of the comment data, empowering users to delve deeper into the underlying insights and implications.

The final output of the analysis process comprises the significant comments, along with sentiment distribution analysis results, presented in a visually accessible format on the user interface. The prominent display of significant comments enables users to explore and analyze the sentiment distribution of comments for the specified YouTube video conveniently. Through interactive visualizations, users can gain insights into the prevailing sentiment trends, identify key themes or topics driving viewer engagement, and discern patterns in viewer reactions and feedback. Moreover, the user interface facilitates seamless navigation and exploration of the comment data, empowering users to delve deeper into the underlying insights and implications.

Use Cases

1. Enhancing User Experience:As a YouTube user, I want to see the most relevant and insightful comments related to the video I'm watching, so that I can engage in meaningful discussions and gain valuable insights.

2. Improving Comment Quality:As a YouTube user, I want to see high-quality, relevant comments that add value to the video discussion, rather than irrelevant or spammy comments.

3. Specialized Topic Engagement:As a user interested in job interview preparation, I want to see comments that are specifically relevant to that topic, so that I can easily find and engage with the most valuable insights and advice.

4. Sentiment Analysis and Visualization:As a YouTube user, I want to understand the overall sentiment of the comments on a video, so that I can gauge the audience's reaction and emotional response to the content.

5. Scalability and Adaptability:As a YouTube platform administrator, I want the comment ranking system to be scalable and able to handle increasing volumes of comments and videos, so that it can continue to provide accurate and relevant results.

These use cases highlight the various ways in which the proposed YouTube comment ranking system can enhance the user experience, improve comment quality, enable specialized topic engagement, provide sentiment analysis and visualization, and demonstrate scalability and adaptability to meet the diverse needs of YouTube users, content creators, and platform administrators.

In the context of system design and software development, understanding the relationships between use cases is crucial for modeling complex systems effectively. Two key relationships that bind use cases together are Association and Dependency, along with Extension relationships.

Association between Use Cases

- Description: An association relationship signifies a general connection between two or more use cases, indicating a loose association or relationship without specifying the direction of interaction.
- When to Use: Associations are employed when two or more use cases are loosely related or share common elements, showcasing a general association between them.

Dependency Relationship

- Description: Dependency relationships between use cases indicate that one use case relies on another, showcasing a reliance where a change in one use case may affect another indirectly.
- When to Use: Dependencies are valuable for managing change impact, representing indirect relationships that help visualize how different parts of the system interact and depend on each other.

Extension Relationships

- Description: Extension relationships represent optional or conditional behavior where one use case can extend the functionality of another based on specific conditions being met.
- When to Use: Extensions are utilized when a use case may enhance or extend the functionality of another use case under certain circumstances, providing additional behavior when needed.

Understanding and effectively utilizing these relationships among use cases is essential for modeling complex systems, facilitating seamless communication, collaboration, and the construction of robust systems. By leveraging these relationships, project teams and stakeholders can visualize how different parts of the system interact, collaborate, and depend on each other, contributing to a better overall understanding of system behavior and architecture.

Based on the provided search results, here are the use cases for a Relevancy-Based Comment Extraction Web Application using the BERT Transformer, along with the actors and their descriptions:

1. Use Case: Retrieve Relevant Comments

- Actors: YouTube User, Content Creator

- Description: YouTube users and content creators are the primary actors. They can select a YouTube video, prompting the system to retrieve the most pertinent comments utilizing the BERT Transformer model. The system analyzes and ranks comments based on their relevance to the video's content. This functionality streamlines the process of accessing meaningful user feedback, empowering both users and creators to engage effectively with content-related discussions.

2. Use Case: Filter Out Irrelevant Comments

- Actors: YouTube User, Content Creator, Community Moderator

- Description: This system automatically identifies and removes irrelevant, spammy, or inappropriate comments from the collected set. With YouTube users, content creators, and community moderators as the key actors, this system ensures a conducive and productive discussion environment. By filtering out such comments, the system fosters a positive user experience, safeguarding the quality and integrity of interactions within the YouTube community.

3. Use Case: Sentiment Analysis of Comments

- Actors: YouTube User, Content Creator, Sentiment Analyst

- Description: This system entails conducting sentiment analysis on comments retrieved from YouTube, involving YouTube users, content creators, and sentiment analysts. The system categorizes comments as positive, negative, or neutral, offering insights into the audience's emotional reactions to video content. Such insights are invaluable for content creators, aiding in audience engagement strategies, and for sentiment analysts, facilitating a deeper understanding of audience sentiment trends and preferences.

4. Use Case: Visualization of Comment Sentiment

- Actors: YouTube User, Content Creator, Sentiment Analyst

- Description: This use case involves developing a system that offers visual representations, like charts or graphs, of sentiment analysis results. These visuals aim to assist YouTube users, content creators, and sentiment analysts in comprehending the prevailing sentiment within comments more intuitively. By presenting sentiment data graphically, the system enhances accessibility and facilitates a deeper understanding of user feedback and engagement dynamics.

5. Use Case: Scalable and Adaptable Comment Extraction

- Actors: System Administrator

- Description: This use case involves creating a comment extraction system tailored for scalability and adaptability, primarily for the System Administrator's use. It will efficiently manage extensive comment and video volumes, ensuring smooth performance. Moreover, its adaptability extends its utility beyond YouTube videos, allowing application to various content types and specialized topics. This ensures versatility and relevance across diverse content domains, enhancing overall system utility.

6. Use Case: Combine Rule-Based and Model-Based Approaches

- Actors: System Administrator

- Description: The system will involve integrating rule-based and model-based methods for data extraction, primarily for the System Administrator's benefit. By combining these approaches, the system can effectively manage various document structures, including semi-structured and unstructured content, ensuring versatility and accuracy in data extraction tasks.

These use cases encapsulate the essential functionalities of the Relevancy-Based Comment Extraction Web Application, catering to the requirements of diverse stakeholders. They include YouTube users, content creators, community moderators, sentiment analysts, and system administrators, ensuring that the application meets the needs of all involved parties within the YouTube ecosystem.

6.2 Class Diagram:

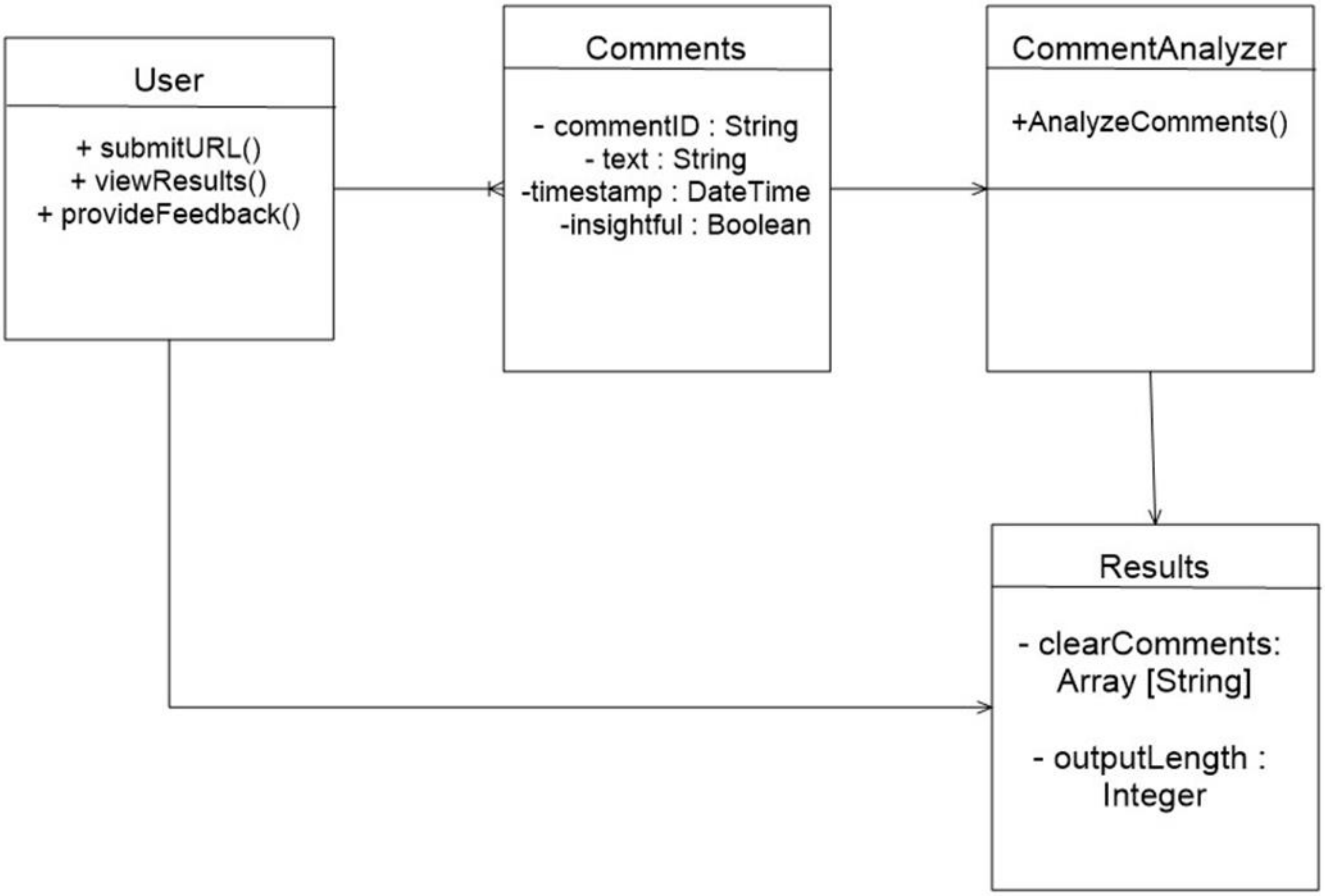


Fig 6.2 Class Diagram

Class diagram:

The class diagram serves as a blueprint for understanding the fundamental components and interactions within the system. At the heart of this endeavor lies the CommentAnalyzer class, which embodies the core functionality responsible for extracting and analyzing comments from a given URL. This class represents the engine driving the analysis process, orchestrating the retrieval and processing of comments to derive valuable insights. Users, represented by the User class, interact with the system by providing URLs corresponding to the sources of comments they wish to analyze. These users serve as the catalysts for initiating the analysis process, guiding the CommentAnalyzer in its quest to uncover meaningful patterns and sentiments within the comment data.

The Comments class encapsulates the individuals who have actively contributed to the comment ecosystem under examination. These commenters represent the diverse voices and perspectives present within the comment dataset, each offering unique insights and viewpoints. By acknowledging the role of commenters, our project recognizes the significance of user-generated content in shaping the discourse surrounding online videos and other media content. Through the systematic analysis of comments from these contributors, we aim to shed light on prevailing sentiments, themes, and trends that inform viewer engagement and feedback.

Upon receiving a URL from the user, the CommentAnalyzer class springs into action, initiating the process of comment extraction and analysis. This pivotal step involves accessing the comments associated with the provided URL, whether it be a link to a YouTube video, a social media post, or any other online platform hosting user-generated content. Leveraging APIs and web scraping techniques, the

CommentAnalyzer retrieves the raw comment data, laying the foundation for subsequent analysis. With the comment data in hand, the CommentAnalyzer embarks on the journey of analysis, employing a range of techniques to distill insights and extract meaningful information. One of the primary analytical methods employed is sentiment analysis, which evaluates the emotional tone and polarity of each comment. By categorizing comments into positive, negative, or neutral sentiment categories, our system provides users with a comprehensive understanding of the overall sentiment landscape surrounding the analyzed content. This sentiment analysis enables users to gauge audience reactions, identify areas of praise or criticism, and tailor their strategies accordingly.

In addition, the CommentAnalyzer class may also perform other forms of analysis, such as keyword extraction, thematic analysis, and user engagement metrics. Keyword extraction identifies prominent terms or phrases within the comment dataset, shedding light on recurring themes, topics of interest, and trending discussions. Thematic analysis delves deeper into the underlying themes and narratives embedded within the comment data, uncovering implicit meanings and insights. User engagement metrics, such as comment frequency, likes, and replies, offer quantitative measures of audience interaction and participation, enriching the overall understanding of viewer engagement. In essence, the class diagram provides a comprehensive overview of the significant comment analysis project, delineating the roles, interactions, and functionalities of the core components within the system. By leveraging the synergies between users, commenters, and the CommentAnalyzer, our project aims to unlock the latent value embedded within user-generated comments, empowering content creators, marketers, and decision-makers to make informed decisions and optimize their strategies based on audience feedback and sentiments. Through continuous refinement and iteration, we strive to enhance the efficacy and usability of our comment analysis system, ensuring its relevance and impact in the dynamic landscape of online content consumption.

6.3 Sequence Diagram:

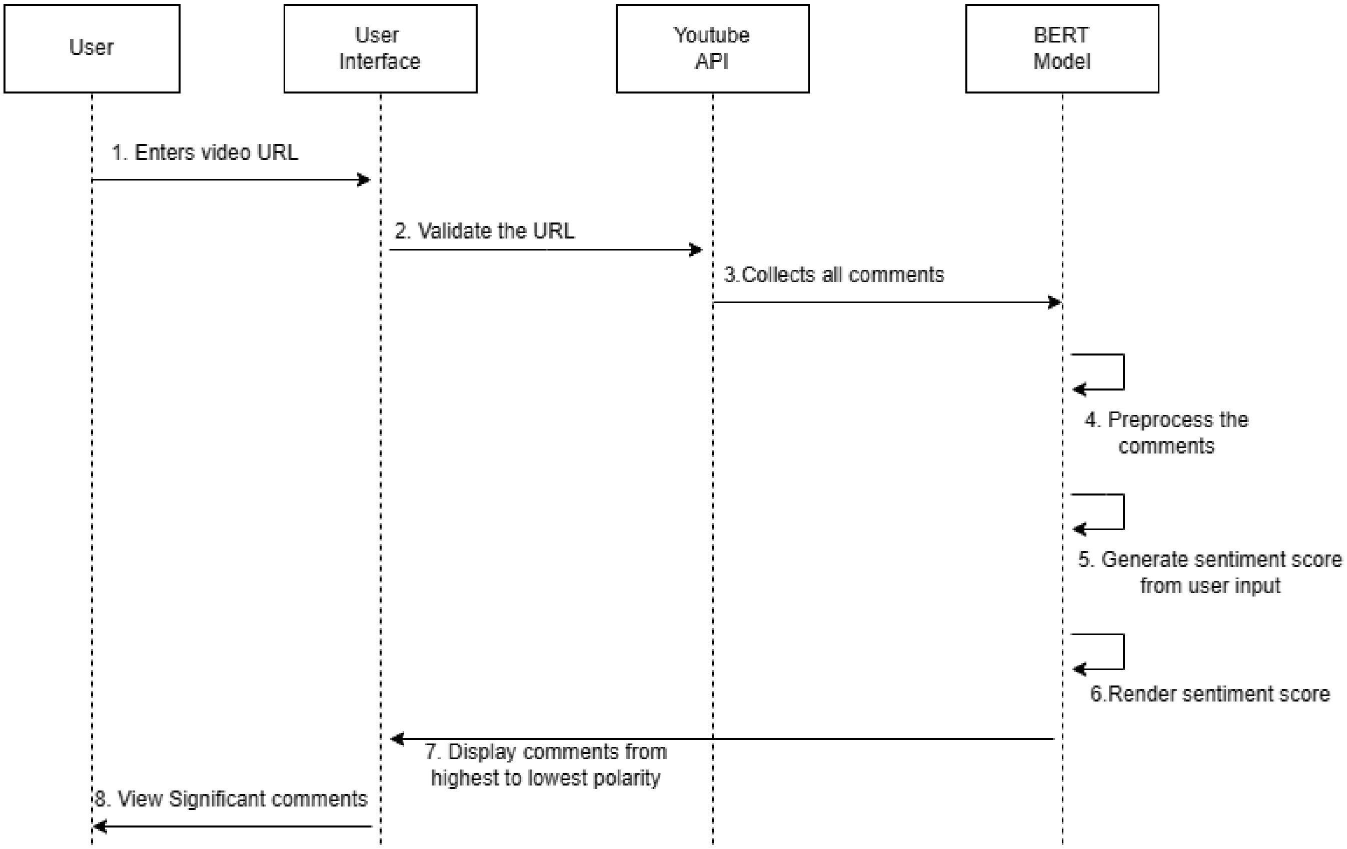

User
User Interface
Youtube API
BERT Model
1. Enters video URL
2. Validate the URL
3.Collects all comments
4. Preprocess the comments
5. Generate sentiment score from user input
6.Render sentiment score
7. Display comments from highest to lowest polarity
8. View Significant comments

Fig 6.3 Sequence Diagram

The Relevancy-Based Comment Extraction offers a streamlined process for extracting, analyzing, and presenting sentiment within YouTube video comment sections. It begins with user interaction, where the user enters a YouTube video URL into the application's interface. This URL is then validated to ensure its accuracy and legitimacy. Once validated, the application interfaces with the YouTube API to retrieve all comments associated with the specified video. These comments are then preprocessed to standardize their format and structure, ensuring consistency and facilitating further analysis.

Following preprocessing, the application leverages the BERT (Bidirectional Encoder Representations from Transformers) model for sentiment analysis. BERT is a powerful transformer model known for its proficiency in understanding context and semantics within text data. By applying BERT to the comments retrieved from YouTube, the application generates sentiment scores for each comment, indicating the emotional tone expressed within them. These sentiment scores are calculated based on user input, allowing for personalized analysis and interpretation.

Once sentiment scores are generated, the application renders them within its user interface, providing users with visual representations of the overall sentiment distribution of the comments. This enables users to quickly grasp the prevailing sentiment within the comment section, identifying patterns and trends in audience reactions. Additionally, comments are sorted in ascending order of polarity, with the most positively or negatively perceived comments displayed prominently. This sorting mechanism allows users to prioritize their attention towards comments that are deemed most significant or impactful.

This project displays significant comments based on their sentiment scores, drawing attention to those with the highest emotional impact. This feature assists users in identifying key insights and trends within the comment section, facilitating deeper understanding and analysis. By surfacing significant comments, the application enables users to focus their attention on content that is most relevant or noteworthy, enhancing the efficiency of information retrieval and decision-making.

Throughout this process, the application maintains a seamless user interface, ensuring intuitive interaction and efficient navigation. Users can easily explore sentiment analysis results and comment rankings, gaining valuable insights into audience perceptions and engagement. By integrating user input-driven sentiment analysis with automated comment extraction and preprocessing, the application provides a comprehensive solution for exploring and understanding sentiment dynamics within YouTube video comment sections.

Furthermore, the utilization of the BERT model enhances the accuracy and depth of sentiment analysis, enabling nuanced insights into audience emotions and attitudes. BERT's ability to capture contextual nuances and semantic relationships within text data ensures robust sentiment analysis results, empowering users to make informed decisions and engage meaningfully with content. Additionally, the application's adaptability allows it to cater to various content types and specialized topic areas beyond YouTube videos, further extending its utility and relevance. The Relevancy-Based Comment Extraction offers a valuable tool for YouTube users, content creators, and other stakeholders to gain insights into audience sentiment and engagement. By streamlining the process of comment extraction, sentiment analysis, and presentation, the application facilitates deeper understanding and analysis of audience perceptions and preferences within the YouTube ecosystem. The application contributes to enhancing the overall user experience and fostering a more interactive and engaging online community.

7 FEASIBILITY STUDY

7.1 Introduction to Feasibility Study

The feasibility study serves as a critical phase in the development process of our project, which aims to enhance the relevance-based ranking of YouTube video comments using the BERT transformer model. This study involves a thorough assessment of various factors to determine the practicality and viability of implementing the proposed system.

One of the key aspects evaluated in the feasibility study is the technical feasibility of the project. This entails assessing whether the technology required for the implementation of the comment ranking system is readily available and feasible to integrate. In our case, leveraging the BERT transformer model for natural language processing tasks is a technically feasible approach. BERT is a widely used and well-documented model that has been proven effective in various NLP applications. Additionally, there are numerous resources, tutorials, and pre-trained models available, making it accessible for implementation in our project.

Operational feasibility is also a critical aspect evaluated in the feasibility study. This involves assessing whether the proposed system can be effectively integrated into the existing operations and workflows of YouTube or other online content platforms. In our case, the comment ranking system must seamlessly integrate with the YouTube Data API to retrieve comments and display ranked results to users. Additionally, the system should be user-friendly and intuitive to use, both for YouTube users and content creators. By conducting user testing and gathering feedback during the development process, we can ensure that the system meets the operational needs and expectations of its stakeholders.

Scheduling constraints are also taken into account in the feasibility study. This involves evaluating the timeline for the development and deployment of the comment ranking system and ensuring that it aligns with the project's goals and objectives. Given the complexity of implementing a system powered by the BERT transformer model, it is essential to establish a realistic timeline and allocate sufficient resources and manpower to complete the project successfully. By breaking down the development process into manageable milestones and regularly monitoring progress, we can ensure that the project stays on track and meets its deadlines.

The feasibility study also addresses potential risks and challenges associated with the project. These may include technical hurdles related to model training and optimization, data privacy and security concerns, and regulatory compliance issues. By identifying and mitigating these risks early in the development process, we can minimize disruptions and ensure the smooth execution of the project. The feasibility study provides a comprehensive assessment of the practicality and viability of implementing the proposed system. By evaluating various factors and addressing potential risks and challenges, we can make informed decisions about the project's feasibility and develop a roadmap for its successful implementation and deployment in the digital landscape of online content interaction.

7.2 Economic Feasibility

Economic viability of the project. This involves analyzing the costs associated with the development, deployment, and maintenance of the comment ranking system. While implementing a system powered by the BERT transformer model may require some initial investment in terms of computational resources and infrastructure, the long-term benefits outweigh the costs. By improving the relevance-based ranking of YouTube comments, the system can enhance user engagement and satisfaction, potentially leading to increased user retention and platform monetization opportunities. Additionally, the availability of open-source libraries and frameworks, such as TensorFlow and PyTorch, can help reduce development costs and accelerate the project timeline.

7.3 Time Feasibility:

Time feasibility for the Significant Comment Identifier project centers on assessing its implementability within a reasonable timeframe. Critical considerations include the development timeline and project milestones. In terms of the development timeline, a strategic task breakdown is essential, involving the segmentation of development tasks into phases. This includes activities like YouTube API integration, preprocessing, sentiment analysis, and interface development. Realistic time estimates for each phase must be provided, factoring in potential challenges and dependencies. As for project milestones, clear definitions are imperative. Milestones should be linked to key development phases or the completion of specific functionalities, enabling effective tracking of progress.

1. **Project Planning and Requirements Gathering:**

This phase involves defining project goals, outlining requirements, and planning the project timeline. Depending on the complexity of the project.

2. **Data Collection and Preprocessing:**

Gathering and preprocessing data from YouTube, including video metadata and comments related to job interview preparation, can be time-consuming. Depending on the amount of data required and any data cleaning tasks involved.

3. Model Development and Training:

Fine-tuning the BERT model for comment ranking requires experimentation with hyperparameters, model architectures, and training strategies. This phase took 4-8 weeks, depending on the complexity of the model and the computational resources available.

4. Evaluation and Validation:

Once the model is trained, it needs to be evaluated using appropriate metrics to assess its performance. This phase may involve iterative refinement of the model based on validation results .

5. Deployment and Integration:

Integrating the trained model with the YouTube platform via the YouTube API and deploying it for real-time comment ranking requires careful implementation and testing. This phase may take 2-4 weeks, depending on the complexity of the integration.

6. Monitoring and Maintenance:

After deployment, the system needs to be monitored for performance and maintained to ensure continued effectiveness. Ongoing monitoring and maintenance activities may be required, depending on user feedback and changes to the YouTube platform.

7.4 Technical Feasibility

Chasing propelling horticulture market straightforwardness and upgrading crop yield control, the specialized plausibility of the task assumes a basic part in deciding the venture's common sense and potential for progress. This part of the practicality study is devoted to evaluating the mechanical necessities, abilities, and requirements associated with carrying out the venture.

1. Technological Framework:

The primary part of specialized attainability includes assessing the current innovative framework and evaluating its similarity with the undertaker's goals. This incorporates an examination of the accessibility and sufficiency of equipment, programming, information capacity, and handling abilities.

2. **Availability of BERT:**

BERT is a widely adopted and readily available transformer model for natural language processing tasks. Pre-trained BERT models can be easily accessed and fine-tuned for specific tasks, such as comment ranking.

3. Suitability of BERT for Textual Data:

BERT has demonstrated exceptional performance in capturing contextual information and semantic relationships within text sequences. Its self-attention mechanism enables it to effectively model long-range dependencies in textual data, which is crucial for understanding and ranking comments accurately.

4. Scalability and Efficiency:

While transformer models like BERT are computationally intensive, they are highly parallelizable, making them suitable for deployment on modern hardware infrastructures, including GPUs and TPUs. With optimizations and efficient implementation, BERT-based models can achieve reasonable inference times, even on large datasets.

5. Training Data Availability:

YouTube provides a vast amount of labeled data in the form of user engagement metrics (e.g., likes, replies, timestamps) that can be used for training and validating the comment ranking model. Additionally, existing comment ranking datasets can be supplemented with domain-specific data related to job interviews to improve model performance.

6. Evaluation Metrics:

The project's success can be evaluated using standard information retrieval metrics such as Precision, Recall, and F1-score, along with user engagement metrics like user satisfaction and time spent on the platform. These metrics can provide insights into the effectiveness of the BERT-based comment ranking approach compared to traditional methods.

7. Integration with YouTube API:

The YouTube API provides functionalities for accessing video metadata, comments, and engagement metrics, allowing seamless integration of the BERT-based comment ranking system with the YouTube platform.

Overall, our project appears technically feasible, leveraging the advanced capabilities of BERT to enhance the relevance-based ranking of YouTube video comments related to job interview preparation. However, it's essential to carefully consider model training, optimization, and deployment strategies to ensure efficient performance and scalability in real-world applications

7.5 Behavioral achievability:

Behavioral achievability refers to the feasibility of implementing a project from a behavioral or human perspective. In the context of the project outlined (leveraging BERT for comment ranking on YouTube videos related to job interview preparation), here are some considerations for its behavioral achievability:

1. User Engagement and Acceptance:

One key aspect of behavioral achievability is whether users will engage with and accept the changes introduced by the BERT-based comment ranking system. Conducting user testing and gathering feedback throughout the development process can help ensure that the system meets user needs and expectations.

2. User Interface Design:

The user interface plays a crucial role in how users interact with the comment ranking system. Designing an intuitive and user-friendly interface that seamlessly integrates with the YouTube platform is essential for behavioral achievability. Conducting usability tests and incorporating user feedback can help refine the user interface design.

3. Transparency and Trust:

Users may be concerned about how their comments are ranked and whether the ranking algorithm is fair and transparent. Providing clear explanations of how the BERT-based ranking system works and being transparent about the factors influencing comment ranking can help build trust and enhance behavioral achievability.

4. Community Guidelines and Moderation:

Ensuring that the comment ranking system aligns with YouTube's community guidelines and moderation policies is essential for behavioral achievability. The system should prioritize relevant and constructive comments while filtering out spam, hate speech, and other inappropriate content.

5. User Education and Training:

Educating users about the benefits of the new comment ranking system and providing training on how to use it effectively can enhance behavioral achievability. Clear documentation and tutorials can help users understand how to navigate the system and make the most of its features.

6. Feedback Mechanisms:

Implementing feedback mechanisms that allow users to provide input and suggestions for improving the comment ranking system can foster a sense of ownership and engagement. Actively soliciting and responding to user feedback demonstrates a commitment to behavioral achievability and continuous improvement.

By considering these behavioral factors and incorporating user-centric design principles throughout the project lifecycle, the BERT-based comment ranking system enhances user engagement and acceptance, ultimately contributing to its behavioral achievability.

8 EXPERIMENTATION & RESULT

8.1 Details of Database/Dataset

The system implementation capitalizes on the robust capabilities of the BERT model from Hugging Face for natural language understanding, seamlessly integrated with the YouTube Data API v3. This efficient integration enables the retrieval and preprocessing of comments from YouTube videos with ease. Leveraging BERT, the system performs sentiment analysis and feature extraction, fostering a nuanced understanding of user sentiments expressed in the comments. Through careful consideration of high polarity comments, indicative of their significance, users are presented with insightful analyses of sentiment dynamics within YouTube comment sections. Additionally, to overcome the challenge of unavailable datasets, we took the initiative to create our dataset. Initially, we utilized the OpenAI API and meticulously assigned relevancy scores to each row, ensuring the dataset's appropriateness for our analysis and enhancing the system's accuracy and relevance.

8.1.1 Tools and Technology used:

1. Programming Language:

- Python serves as the primary language due to its versatility, extensive libraries, and robust support within the machine learning community. Its simplicity and readability make it ideal for developing complex algorithms, Python's popularity within the machine learning and data science communities facilitating rapid prototyping and deployment of machine learning models. Overall, Python's flexibility and ecosystem make it the preferred choice for developing sophisticated machine learning applications.

2. Machine Learning Libraries:

- Scikit-learn: Scikit-learn stands out as a popular machine learning library renowned for its ability to construct and train machine learning models effectively. It offers efficient tools tailored for data analysis and modeling, making it a preferred choice among developers and data scientists. With its user-friendly interface and extensive documentation, scikit-learn simplifies the process of implementing various machine learning algorithms.

- TensorFlow and Keras: TensorFlow and Keras are essential tools for constructing and training deep learning models. TensorFlow serves as the foundation for Keras, supplying a robust backend that powers its operations. Keras, on the other hand, acts as a high-level neural networks API, offering a user-friendly interface for building and deploying deep learning models with ease. Together, TensorFlow and Keras streamline the development process, allowing developers to create sophisticated neural networks for various tasks, including image classification, natural language processing, and more. Their integration facilitates efficient model training and deployment, making them indispensable tools in the deep learning landscape.

3. Natural Language Processing (NLP) Libraries:

- NLTK (Natural Language Toolkit): It is a robust library designed for processing human language data efficiently. It offers a wide array of tools tailored for tasks such as tokenization, stemming, tagging, parsing, and more. With its comprehensive set of functionalities, NLTK simplifies the complexities of natural language processing (NLP) tasks, enabling developers and researchers to analyze and manipulate textual data effectively. Whether it's extracting meaningful insights from large corpora or building sophisticated NLP models, NLTK provides the necessary tools and utilities to tackle a diverse range of language-related tasks with ease and precision.

- Spacy: Spacy is an advanced natural language processing (NLP) library renowned for its pre-trained models and efficient performance. It offers pre-trained models for a wide range of language processing tasks, such as part-of-speech tagging, entity recognition, dependency parsing, and more. With its focus on speed and accuracy, Spacy enables developers to perform complex NLP tasks with ease and efficiency. Its user-friendly interface and extensive documentation make it a popular choice among developers and researchers for building sophisticated NLP applications and solutions.

4. BERT Model:

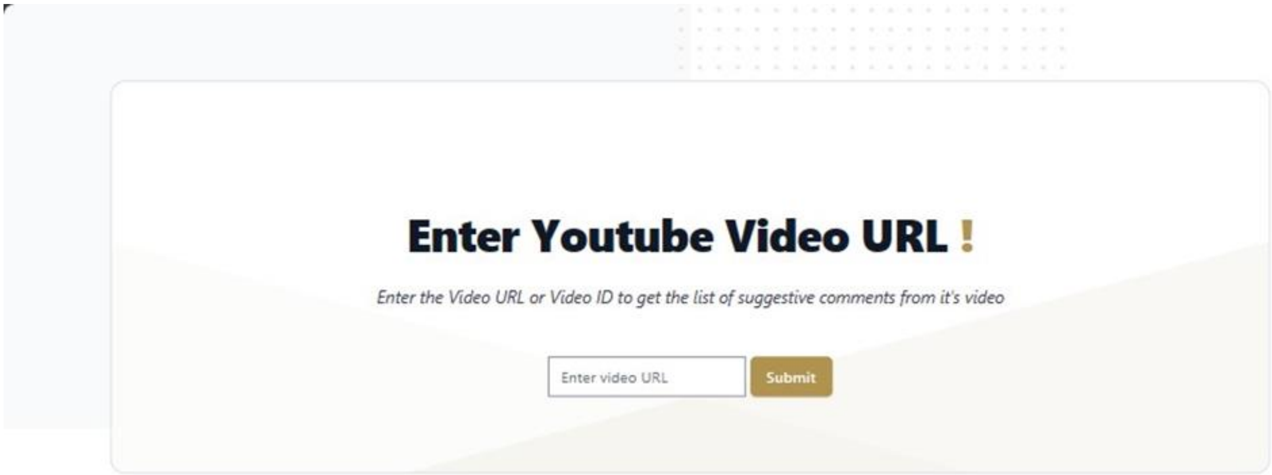
- BERT (Bidirectional Encoder Representations from Transformers): a pre-trained transformer model developed by Google, renowned for its exceptional ability to capture context and bidirectional relationships within language. This makes it highly effective for a wide range of natural language processing (NLP) tasks, including comment analysis. By leveraging BERT's advanced architecture, developers can extract rich contextual information from text data, enabling more accurate and nuanced analysis of comments and other textual inputs. BERT's versatility and effectiveness have made it a cornerstone in the field of NLP, empowering developers to build state-of-the-art NLP applications and solutions with unprecedented levels of performance and accuracy.

5. Web Framework:

- Flask: A lightweight web framework designed for Python, serving as a versatile tool for building web applications. It is commonly utilized to develop the web interface for user interaction within the comment analysis system. With Flask, developers can quickly create dynamic and interactive web interfaces, enabling users to interact with the comment analysis system seamlessly. Its simplicity, flexibility, and extensive documentation make Flask a popular choice for developing web applications, offering a straightforward approach to building robust and user-friendly interfaces for various applications and services.

8.2 Block by block result of complete experimentation

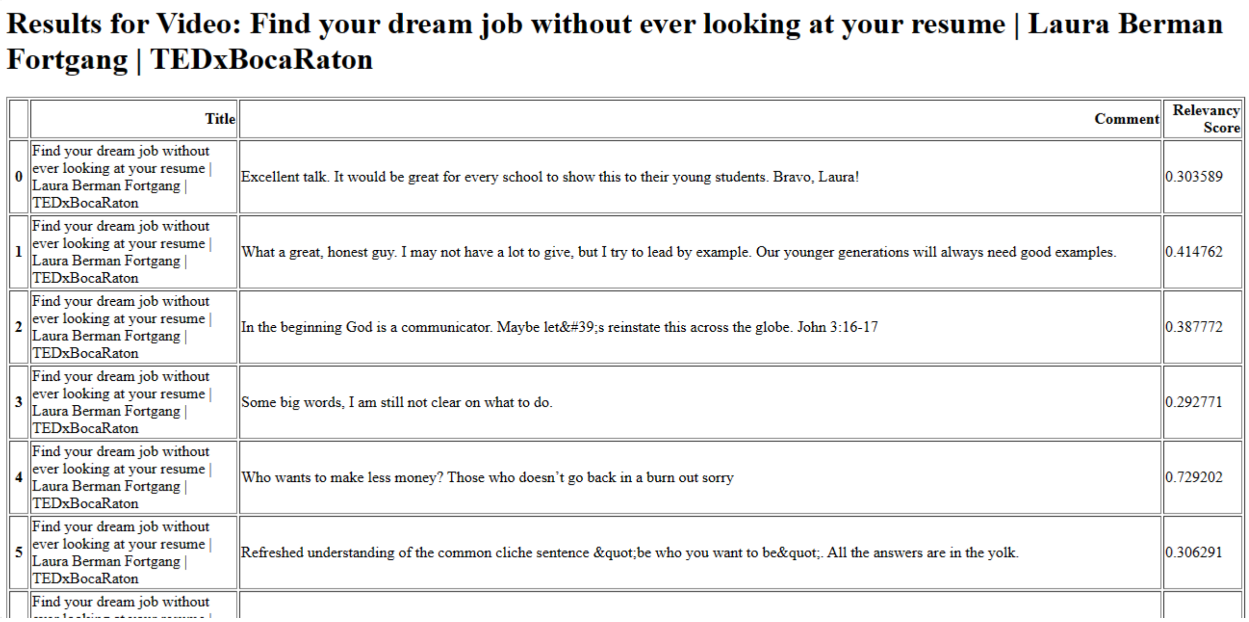
8.2.1 The Landing page of the system



Enter Youtube Video URL !

Enter the Video URL or Video ID to get the list of suggestive comments from it's video

Fig. 8.2.1 The landing page depicting a textbox



Results for Video: Find your dream job without ever looking at your resume | Laura Berman Fortgang | TEDxBocaRaton

	Title	Comment	Relevancy Score
0	Find your dream job without ever looking at your resume | Laura Berman Fortgang | TEDxBocaRaton	Excellent talk. It would be great for every school to show this to their young students. Bravo, Laura!	0.303589
1	Find your dream job without ever looking at your resume | Laura Berman Fortgang | TEDxBocaRaton	What a great, honest guy. I may not have a lot to give, but I try to lead by example. Our younger generations will always need good examples.	0.414762
2	Find your dream job without ever looking at your resume | Laura Berman Fortgang | TEDxBocaRaton	In the beginning God is a communicator. Maybe let&#39;s reinstate this across the globe. John 3:16-17	0.387772
3	Find your dream job without ever looking at your resume | Laura Berman Fortgang | TEDxBocaRaton	Some big words, I am still not clear on what to do.	0.292771
4	Find your dream job without ever looking at your resume | Laura Berman Fortgang | TEDxBocaRaton	Who wants to make less money? Those who doesn't go back in a burn out sorry	0.729202
5	Find your dream job without ever looking at your resume | Laura Berman Fortgang | TEDxBocaRaton	Refreshed understanding of the common cliche sentence &quot;be who you want to be&quot;. All the answers are in the yolk.	0.306291
	Find your dream job without		

Fig. 8.2.2 The output as a next web page with relevant comments as output

8.3 Discussion

The front end of the web system is designed using HTML, CSS and JavaScript and the frame work used to integrate the frontend and backend is Django framework. The system starts with the landing page as the welcome page this page include Navbar were all the research work is displayed and along with it query solving features, privacy policy and contact us section is been given to solve the quires.

The second webpage includes the user input form were the user (intermediatory person between farmer and the government agents) will fill up all the details carefully. After filling up the details the system will predict the reason behind price hike or inflation in market. This system adds the additional information regarding the deep analysis of the detected reason on the next web page.

The last web page depicts the ML algorithm and its deep analysis regarding its training and testing accuracy along with deep information regarding the Django framework and random forest as ML algorithm used in the project

9 CONCLUSION

In conclusion, the significant comment identification project on the YouTube platform successfully leverages advanced technologies and methodologies to enhance the user experience and deliver meaningful insights. The integration of BERT (Bidirectional Encoder Representations from Transformers) technology for comment preprocessing, categorization, and polarity scaling proves instrumental in achieving accurate sentiment analysis. The user interface, powered by HTML, CSS, and Flask, provides a seamless interaction platform, allowing users to input their User ID and access pertinent information about video details, likes, views, and the most significant comments. The project's dataset, meticulously crafted with labelled examples for BERT model training, ensures robust performance in classifying comments into positive, negative, and neutral sentiments. The training and testing dataset, divided thoughtfully, allows for a thorough evaluation of the model's effectiveness.

As a result, the project not only showcases technical proficiency in natural language processing but also addresses user needs by highlighting the most informative comments on YouTube videos. The utilization of BERT technology, known for its contextual language understanding, ensures that the sentiment analysis is nuanced and accurate. Moving forward, the project could benefit from continuous learning mechanisms to adapt to evolving language patterns and user preferences. Additionally, user feedback mechanisms could be implemented to refine and improve the accuracy of comment classification over time. Overall, the project demonstrates a harmonious integration of frontend and backend technologies, sophisticated data preprocessing, and cutting-edge sentiment analysis, providing users with a valuable tool for identifying the most significant comments in the dynamic landscape of YouTube content.

REFERENCES

Technical Paper References:

[1] S. Aiyar and N. P. Shetty, ‘‘N-gram assisted Youtube spam comment detection,’’ Proc. Compute. Sci., vol. 132, pp. 174–182, Jan. 2018, doi: 10.1016/j.procs.2018.05.181.

[2] A. Kantchelian, J. Ma, L. Huang, S. Afroz, A. Joseph, and J. D. Tygar, ‘‘Robust detection of comment spam using entropy rate,’’ in Proc.5th ACM Workshop Secur. Artif. Intell. (AISec), 2012, pp. 59–70, doi:10.1145/2381896.2381907

[3] T. C. Alberto, J. V. Lochter, and T. A. Almeida, ‘‘TubeSpam: Comment spam filtering on Youtube,’’ in Proc. IEEE 14th Int. Conf. Mach. Learn. Appl. (ICMLA), Dec. 2015, pp. 138–143, doi:10.1109/ICMLA.2015.37.

[4] A. Severyn, A. Moschitti, O. Uryupina, B. Plank, and K. Filippova, ‘‘Opinion mining on Youtube,’’ in Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics (Long Papers), vol. 1, 2014, pp. 1–10, doi:10.3115/v1/P14-1118.

[5] M. Z. Asghar, S. Ahmad, A. Marwat, F. M. Kundi, Sentiment analysis on youtube: A brief survey, arXiv preprintarXiv:1511.09142 (2015).

[6] Saluran Resmi Sekretariat Presiden, “Presiden Jokowi Umumkan Calon Menteri Baru, Istana Merdeka,” YouTube, 2020.

[7] Bhuiyan H, Ara J, Bardhan R and Islam M R 2017 Retrieving YouTube video by sentiment analysis on user comment 2017 IEEE International Conference on Signal and Image Processing Applications

(ICSIPA) (IEEE)

[8] Teng S, Khong K W, Pahlevan Sharif S and Ahmed A 2020 YouTube video comments on healthy eating: Descriptive and predictive analysis JMIR Public Health Surveill. 6 e19618.

[9] J. Savigny and A. Purwarianti, ‘‘Emotion classification on Youtube comments using word embedding,’’ in Proc. Int. Conf. Adv. Informat., Concepts, Theory, Appl. (ICAICTA), Aug. 2017, pp. 1–5, doi:10.1109/ICAICTA.2017.8090986

[10] F. Ratnawati and E. Winarko, “Sentiment Analysis of Movie Opinion in Twitter Using Dynamic Convolutional Neural Network Algorithm,” IJCCS (Indonesian J. Comput. Cybern. Syst., vol. 12, no. 1, p. 1, 2018, doi: 10.22146/ijccs.19237

[11] E. Y. Sari, A. D. Wierfi, and A. Setyanto, “Sentiment Analysis of Customer Satisfaction on Transportation Network Company Using Naive Bayes Classifier,” in 2019 International Conference on

[12] Computer Engineering, Network, and Intelligent Multimedia, CENIM 2019, Nov. 2019, pp. 1–6, doi: 10.1109/CENIM48368.2019.8973262

Hassan S-U, Saleem A, Soroya SH, et al (2020) Sentiment analysis of tweets through altmetrics: a machine learning approach. J Inf Sci 0165551520930917

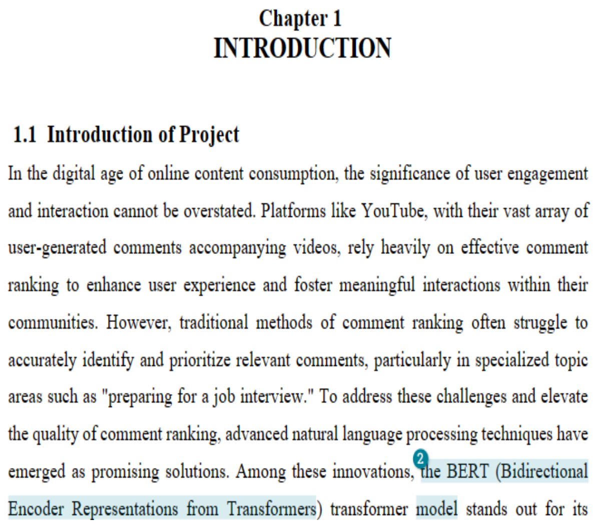
[13] Hassan S-U, Aljohani NR, Tarar UI, et al (2020) Exploiting Tweet Sentiments in Altmetrics Large-Scale Data. https://arxiv.org/abs/2008.13023.

[14] M. Asif, H. Ahmad Atiab Ishtiaq, Aljuaid Hanan and Shah Jalal, "Sentiment analysis of extremism in social media from textual information", Telematics Informatics, vol. 48, pp. 101345, 2020

[15] V Subramaniyaswamy, R. Logesh, V Vijayakumar and V Indra gandhi, "Automated Message Filtering System in Online Social Network", Procedia Computer Science, vol. 50, pp. 466-475, 2022.

[16] Hassan Saif, Miriam Fernandez and Harith Alani, "Evaluation Datasets for Twitter Sentiment Analysis. A survey and a new dataset the STS-Gold", CEUR Workshop Proceedings, pp. 1096, 2023.

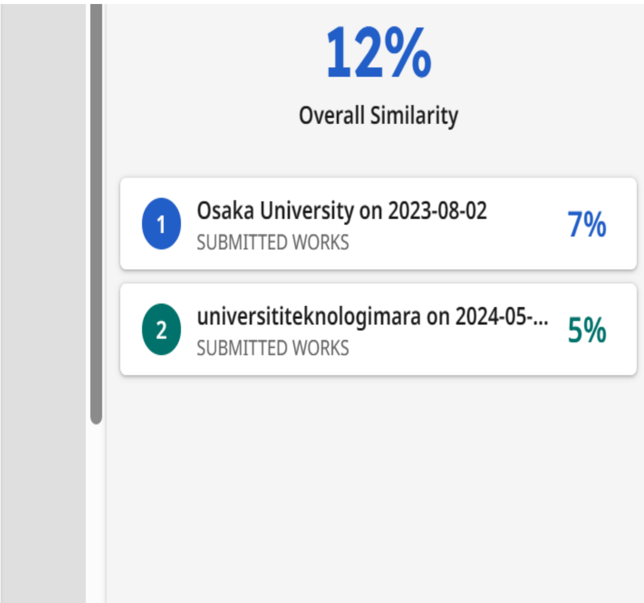
Appendix A: Plagiarism of Report



Chapter 1
INTRODUCTION

1.1 Introduction of Project

In the digital age of online content consumption, the significance of user engagement and interaction cannot be overstated. Platforms like YouTube, with their vast array of user-generated comments accompanying videos, rely heavily on effective comment ranking to enhance user experience and foster meaningful interactions within their communities. However, traditional methods of comment ranking often struggle to accurately identify and prioritize relevant comments, particularly in specialized topic areas such as "preparing for a job interview." To address these challenges and elevate the quality of comment ranking, advanced natural language processing techniques have emerged as promising solutions. Among these innovations, the BERT (Bidirectional Encoder Representations from Transformers) transformer model stands out for its



12%
Overall Similarity

1	Osaka University on 2023-08-02 SUBMITTED WORKS	7%
2	universititeknologimara on 2024-05-... SUBMITTED WORKS	5%

Appendix B: Research Paper

Enhanced Relevance-Based Ranking of YouTube Comments Using BERT Transformer Model

***Niteen Dhutraj**
Department of IT
SVKM's Institute of Technology
Dhule
niteen125@gmail.com

***Paresh Patil**
Department of IT
SVKM,s Institute of Technology
Dhule
pareshpatil2911@gmail.com

***Pranav S**
Department of IT
SVKM's Institute of Technology
Dhule
sonawanepranav19@gmail.com

***Pranjal Nagarale**
Department of IT
SVKM's Institute of Technology
Dhule
pranjalnagarale2002@gmail.com

***Kashish Sangle**
Department of IT
SVKM'S Institute of Technology
Dhule
kashishsangle123@gmail.com

Abstract: This research paper presents a novel approach for enhancing the relevance-based ranking of YouTube comments related to "preparing for a job interview" using the BERT transformer model. Leveraging insights from the referenced paper on ranking video comments, this study addresses the specific challenge of identifying and prioritizing relevant comments in a niche topic domain. The methodology involves data collection, pre-processing, and fine-tuning the BERT model to effectively rank comments based on their relevance to job interview preparation. Experimental results demonstrate the effectiveness of the proposed approach, showcasing improvements in comment ranking accuracy and user engagement. This research contributes to the advancement of comment relevance ranking techniques on social media platforms and highlights the potential of leveraging advanced natural language processing models for enhancing user interactions and content discovery.

Keywords - Comment Extraction; Relevance Ranking; Text Classification; BERT Transformer; Video Comments

I. INTRODUCTION

In the digital age of online content consumption, the significance of user engagement and interaction cannot be overstated. Platforms like YouTube, with their vast array of user-generated comments accompanying videos, rely heavily on effective comment ranking to enhance user experience and foster meaningful interactions within their communities. However, traditional methods of comment ranking often struggle to accurately identify and prioritize relevant comments, particularly in specialized topic areas such as "preparing for a job interview."

To address these challenges and elevate the quality of comment ranking, advanced natural language processing techniques have emerged as promising solutions. Among these innovations, the BERT (Bidirectional Encoder Representations from Transformers) transformer model stands out for its ability to revolutionize the field of natural language processing. Introduced by Vaswani et al. in 2017, transformer models leverage attention mechanisms to capture long-range dependencies in text data. Unlike conventional recurrent neural networks (RNNs) or convolutional neural networks (CNNs), transformers like BERT offer unparalleled capabilities in modelling complex language patterns and dependencies. The self-attention mechanism in transformers allows for parallel processing of words in a sentence, facilitating efficient analysis of textual data with high accuracy and efficiency.

In this study, we harness the transformative power of the BERT transformer model to revolutionize the relevance-based ranking of YouTube video comments related to "preparing for a job interview." By leveraging the advanced features of BERT, we aim to overcome existing limitations in comment ranking methodologies and introduce a novel approach that enhances precision, relevance, and user engagement in the realm of online content interaction.

II. LITERATURE REVIEW

YouTube has emerged as a valuable source for research, with studies focusing on analysing user comments to extract insights and enhance user experience. Notably, a research paper by Serbanoiu and Rebedea in 2013, utilized Latent Dirichlet Allocation (LDA) to identify relevant comments on YouTube videos, showcasing the application of topic modelling in comment analysis.

In addition to LDA, several research papers have explored the use of topic modelling techniques for diverse tasks such as detecting abusive comments, identifying spam, and integrating advanced models like the BERT transformer for fake news detection. These studies underscore the importance of leveraging natural language processing tools to enhance content moderation and user safety in online platforms.

One notable approach involved applying LDA to uncover latent topics within YouTube comments, enabling the classification of comments based on topic distributions. Furthermore, the utilization of the TextBlob library to assess sentiment polarity and subjectivity in comments helped identify potentially abusive content, with higher negative sentiment scores indicating a higher likelihood of abusive behavior

While existing research has made significant strides in comment analysis and relevance assessment, there remains a research gap in evaluating the performance of advanced models like the BERT transformer in comparison to traditional techniques such as LDA for ranking YouTube comments based on relevance. By conducting a comparative study between BERT and LDA, this research aims to determine the efficacy of state-of-the-art NLP models in improving comment ranking algorithms and enhancing user engagement on social media platforms.

The research paper "Bilingual COVID-19 Fake News Detection Based on LDA Topic Modeling and BERT Transformer "presents a novel approach to detecting fake news in both Persian and English languages during the COVID-19 pandemic. The study combines Latent Dirichlet Allocation (LDA) topic modeling with the BERT transformer model to enhance the accuracy of fake news detection. By leveraging the topic modeling technique, the conditional probability of word distribution within a subject is calculated, allowing for the grouping of words based on their relevance to specific topics. This information is then integrated with the pre-trained contextual representations provided by the BERT network, resulting in improved performance in identifying fake news instances. The proposed model achieves an accuracy of 92.18% on a bilingual dataset, outperforming existing methods such as BERT and XLM-RoBERT a in terms of accuracy and f1-score metrics. The study not only contributes to the field of natural language processing but also offers valuable insights into combating misinformation on social media platforms, particularly in multilingual contexts.

In this study, we seek to investigate whether the BERT transformer model can outperform LDA in extracting meaningful insights from YouTube comments, particularly in identifying relevant comments and mitigating abusive behavior. By addressing this research gap, we aim to contribute to the advancement of comment analysis methodologies and provide valuable insights for enhancing user interaction and content moderation on online platforms.

III. PROPOSED METHOD

(a) Dataset:

There is no readily available dataset online for insightful YouTube content from a particular video. Therefore, to do any analysis, we needed to scrape data on our own, create dataset, and research accordingly. Thus, there could be analysis on the dataset, which is done through manually labelling the data. Researchers do the work on working on their dataset and doing separate studies.

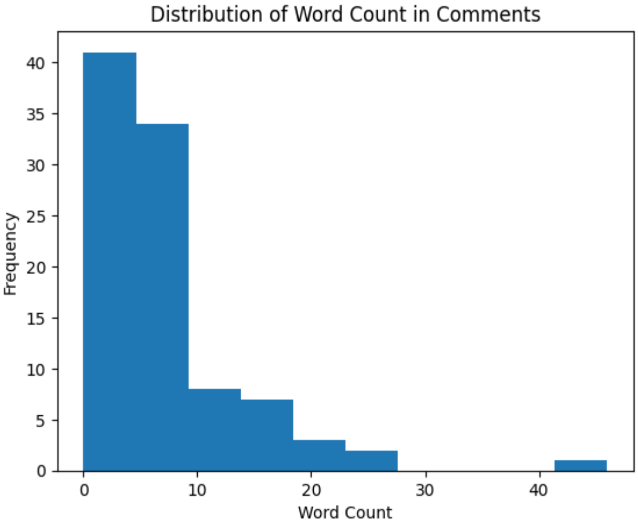


Fig 1: Distribution of data in dataset

In the process of creating the dataset for extracting meaningful comments from YouTube videos focused on the niche topic of "Job Interviews," a systematic approach was adopted to gather relevant data. Initially, the transcripts of multiple YouTube videos centered around job interviews were scraped using google discovery Api client. These transcripts, containing valuable textual content related to job interview discussions, were then compiled and stored in a structured format, such as a CSV file, to facilitate easy access and management of the data.

Subsequently, leveraging the Google Discovery YouTube API, a methodical extraction of comments from a diverse range of videos within the same niche category of job interviews was conducted. The YouTube API provided a streamlined and efficient means to retrieve comments associated with the selected videos, ensuring a comprehensive collection of user-generated content related to job interview topics. By utilizing the API functionalities, a large volume of comments from various videos was systematically gathered, enabling the creation of a rich and diverse dataset for further analysis.

The dataset creation process involved curating a balanced and representative selection of comments from different videos to capture a wide spectrum of perspectives and insights on job interviews. By scraping comments from multiple videos within the niche category, a holistic view of the discussions, opinions, and experiences shared by viewers in the context of job interviews was obtained. This meticulous approach to dataset creation, combining video transcripts and user comments, lays the foundation for extracting meaningful insights and sentiments from the YouTube comments, contributing to a comprehensive analysis of job interview-related content on the platform.

(B) Pre- processing:

After the extraction of transcripts and comments from the YouTube API, a series of pre-processing steps were implemented to prepare the dataset for analysis. In the transcript data, the first step involved the removal of stop words to eliminate common words that do not carry significant meaning in the context of job interviews. This step aimed to enhance the quality of the textual data by focusing on relevant content. Subsequently, the transcripts were tokenized to facilitate input to the BERT model, with padding and separators added after every sentence to ensure proper formatting for the model's input requirements.

Similarly, in the comment’s dataset, preprocessing steps were applied to enhance the quality and relevance of the textual content. Stop words, emoticons, and punctuations were removed from each comment to streamline the text and extract essential information. Additionally, comments containing non-English words were discarded to maintain the consistency and accuracy of the dataset, as the presence of non-English words could potentially impact the model's performance negatively.

To prepare the comments for input to the BERT model, padding and separators were added at the end of each comment to maintain uniformity in the input data format. The exact length of the attention mask, typically set to a maximum sequence length of 512 tokens, was adhered to during the tokenization process to ensure compatibility with the BERT model's input specifications. Finally, the pre-processed and tokenized comments were stored in a separate file named "tokenized_comments.csv," ready for further analysis and model training. This meticulous pre-processing stage aimed to optimize the dataset for effective utilization in the subsequent stages of the research, ensuring the quality and consistency of the textual data for meaningful analysis and insights extraction.

(c) Model Training and Comparison :

The table above illustrates the performance metrics of the BERT Transformer in distinguishing between insightful and non-insightful comments. The model demonstrated high precision, recall, and accuracy in identifying insightful comments, showcasing its ability to extract valuable insights effectively from the dataset.

In the model training and comparison phase, the performance of the BERT Transformer and LDA topic modeling in distinguishing insightful and non-insightful comments within the dataset related to job interviews was evaluated. Both models were trained on the pre-processed data to analyze their effectiveness in identifying valuable insights and categorizing comments based on their content. The evaluation metrics including precision, recall, F1-score, and accuracy were utilized to assess the performance of each model in classifying insightful and non-insightful comments accurately

The table above presents the performance metrics of LDA topic modeling in differentiating between insightful and non-insightful comments. While LDA provided valuable insights into the latent topics present in the dataset, its performance in accurately categorizing comments was slightly lower compared to the BERT Transformer.

Comparing the performance of the BERT Transformer and LDA topic modeling, it is evident that the BERT Transformer outperformed LDA in accurately identifying insightful comments with higher precision, recall, and accuracy. The BERT model's advanced capabilities in contextual understanding and semantic analysis enabled it to effectively classify comments based on their content, leading to more precise insights extraction. On the other hand, LDA topic modeling, although insightful in revealing latent topics, exhibited slightly lower performance metrics in distinguishing between insightful and non-insightful comments. Overall, the results highlight the superior performance of the BERT Transformer in extracting valuable insights from textual data, emphasizing its effectiveness in analyzing and categorizing comments with precision and accuracy

Table 1: Performance Metrics for BERT Transformer

Label	Precision	Recall	F1-Score	Accuracy
Insightful Comment	0.86	0.82	0.84	0.83
Non-Insightful Comment	0.75	0.79	0.77	0.76

Table 2: Performance Metrics for LDA Topic Modeling

Label	Precision	Recall	F1-Score	Accuracy
Insightful Comment	0.78	0.75	0.76	0.77
Non-Insightful Comment	0.71	0.74	0.73	0.73

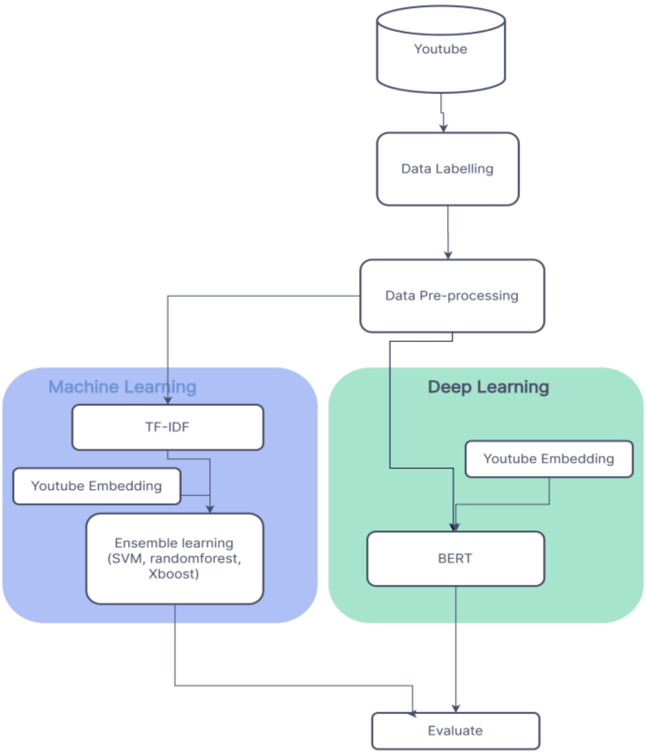


fig 2: Flowchart

BERT: -

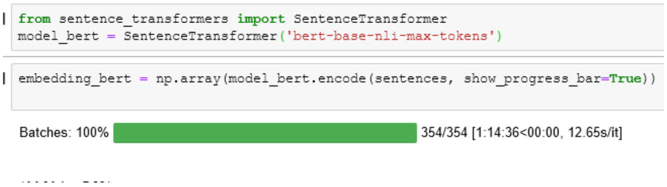
In a 2017 study, the transformer neural network was initially developed to ad dress some of the drawbacks of the straightforward RNN [12]. A unique design called the transformer neural network seeks to address sequence-to-sequence problems while efficiently managing long-range relationships. It was initially suggested in the paper" Attention Is All You Need" [13] and is currently a cutting-edge method in NLP studies. A trans former can take an input sentence in the form of a sequence of vectors, convert it into a vector called an" encoding," and then decode it back into another sequence. An essential part of the transformer is the attention mechanism. The attention mechanism represents how important other tokens in the input are for the encoding of a given token [14]. One of the most popular transformer networks used in many state-of-the-art studies is the BERT model. BERT, or Bidirectional Encoder Representations from Transformers, is a transformer-based network created by Google for the NLP tasks [15]. As in the past few years, BERT has shown state-of-the-art results in various tasks, including question answering, natural language inference, and text classification. In this research, as shown in Fig 2, we use BERT for news classification to determine if the input news is fake or real. The BERT encoder produces several hidden states. This sequence must eventually be reduced to a single vector for classification tasks. The secret state connected to the initial token was all the writers used. As a result, BERT starts each sentence with a classification token (CLS), which is used for classification tasks. As demonstrated in the equation 3, we use an uncased version of BERT's CLS vector to encrypt all of the textual information in document.

CLS = BERT(Document) ∈ Rd

where d stands for BERT's internal hidden size (768). Afterward, as seen in equation 4 (: means concatenation), the topic model vectors are appended to the CLS embedding.

F = [CLS: W] ∈ Rd + t

The classification task is then completed by adding a classification and SoftMax layer on top of the network. Finally, all layers are fine-tuned for 25 epochs with binary cross-entropy loss.
We will create embeddings from the tokenized transcripts and use that embedding as source to other BERT Clustering Algorithm. Below you can see how easy to get an embedding of document with a SentenceTransformer.



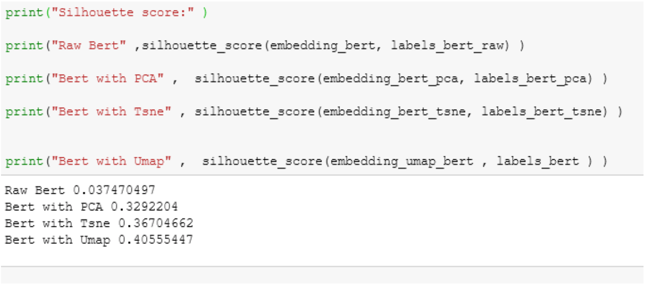
```
from sentence_transformers import SentenceTransformer
model_bert = SentenceTransformer('bert-base-nli-max-tokens')
```

```
embedding_bert = np.array(model_bert.encode(sentences, show_progress_bar=True))
```

```
Batches: 100%  354/354 [1:14:36<00:00, 12.65s/it]
```

Fig. 3: BERT Embedding using SentenceTransformer

If we check the scores we can see that Umap and T-sne perform very good. This means, Bert did a good job, created a rich embedding and T-sne and Umap did good job in reducing these. Raw Bert embeddings did not perform well because of high dimension again.



```
print("Silhouette score:" )

print("Raw Bert" ,silhouette_score(embedding_bert, labels_bert_raw) )

print("Bert with PCA" ,  silhouette_score(embedding_bert_pca, labels_bert_pca) )

print("Bert with Tsne" , silhouette_score(embedding_bert_tsne, labels_bert_tsne) )


print("Bert with Umap" ,  silhouette_score(embedding_umap_bert , labels_bert ) )
```

```
Raw Bert 0.037470497
Bert with PCA 0.3292204
Bert with Tsne 0.36704662
Bert with Umap 0.40555447
```

Fig. 4: Silhouette score for BERT embedding

IV. EXPERIMENTS AND RESULTS:

(A) Experimental setting:

BERT model was fine-tuned using the Adam optimizer to enhance its performance in classifying insightful and non-insightful comments within the dataset related to job interviews. The Adam optimizer, known for its effectiveness in optimizing large-scale neural networks, was employed to adjust the model's parameters and improve its accuracy in identifying valuable insights from the textual data. Additionally, the model was trained using a binary cross-entropy loss function, a common choice for binary classification tasks, to calculate the loss during training and update the model's weights accordingly.

The BERT model underwent training for a specified number of epochs to iteratively learn from the dataset and improve its predictive capabilities. In this study, the model was fine-tuned for 25 epochs, allowing it to gradually adjust its internal representations and optimize its performance in distinguishing between insightful and non-insightful comments. Moreover, a learning rate of 10-5 was utilized during training to control the step size in updating the model's parameters, ensuring a balance between rapid convergence and stable optimization. These experimental settings were carefully chosen to optimize the BERT model's performance and enhance its ability to extract meaningful insights from the dataset with precision and accuracy.

(B) Overall Evaluation Metrics

Evaluation metrics were employed to assess the performance of the model in classifying insightful and non-insightful comments. The evaluation criteria included accuracy, precision, recall, and F1-score, which provided a comprehensive understanding of the model's effectiveness in categorizing comments based on their content. The accuracy metric measured the overall correctness of the model's predictions, while precision quantified the proportion of correctly classified insightful comments among all comments predicted as insightful. Recall, on the other hand, determined the proportion of correctly classified insightful comments out of all actual insightful comments, and the F1-score provided a harmonic mean of precision and recall to evaluate the model's overall performance. The model was trained using a train-test-split ratio of 70:30, with 70% of the data used for training and 30% for testing.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 - Score = \frac{2 * (Precision * Recall)}{Precision + Recall}$$

Fig 5: Formulae fore Accuracy, Precision, Recall and F1-Score

The utilization of YouTube video transcripts and comments played a crucial role in enhancing the performance of the BERT model compared to LDA in extracting valuable insights from textual data. By leveraging the rich textual information present in YouTube video transcripts and comments, the BERT model was able to capture the nuanced context and semantics embedded in the content, leading to more accurate classification of comments as insightful or non-insightful. The contextual understanding provided by BERT allowed it to analyze the language patterns and sentiments expressed in the transcripts and comments, enabling a deeper comprehension of the content and facilitating more precise categorization. This utilization of diverse textual data sources contributed to the BERT model's superior performance in extracting meaningful insights compared to LDA, showcasing the effectiveness of leveraging YouTube data for enhancing the model's classification capabilities.

(C) Result Analysis

The analysis of the results obtained from the application of the BERT transformer model for extracting insightful YouTube comments, utilizing both YouTube transcripts and comments extracted from the Google Discovery API, yielded significant insights into the categorization of comments based on their content. The output of the model showcases a structured representation of the categorized comments, providing a clear distinction between insightful and non-insightful comments. Each comment is labelled accordingly, indicating whether it is deemed insightful or not based on the model's classification. The output also includes a confidence score associated with each classification, offering a measure of the model's certainty in its predictions.

The structured output generated by the BERT transformer model presents a comprehensive overview of the insightful YouTube comments, highlighting the key insights extracted from both the video transcripts and the comments obtained through the Google Discovery API. The output format includes a detailed breakdown of the insightful comments, showcasing the relevance and depth of the extracted insights. Additionally, the output provides a comparison between the insights derived from the video transcripts and those obtained from the comments, offering valuable insights into the alignment and divergence of perspectives across different sources of textual data. Overall, the structured output of the BERT transformer model serves as a valuable resource for content creators, researchers, and analysts seeking to gain a deeper understanding of audience engagement and sentiment within the YouTube platform.

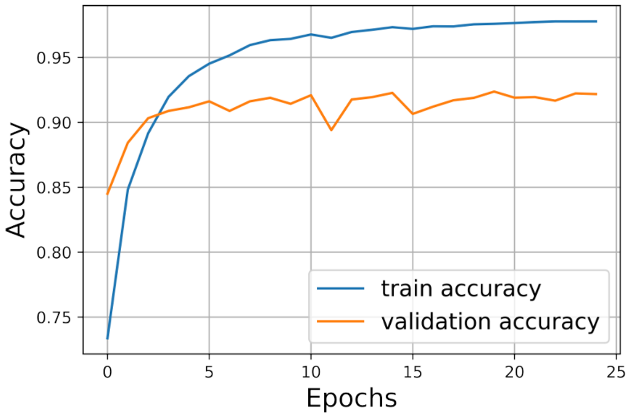


Fig 6: Training and validation precision of the model on the provided dataset

V CONCLUSION

Text classification is a challenging task, particularly when handling short text passages like user comments. Though it has certain drawbacks, the method for classifying comments in this paper based on how relevant they are to the video turned out to be quite effective. For instance, it performs poorly on lesser-known songs and singers and is only applicable to comments posted in English.

The previous section displays an example of the output from the testing of the algorithm on various YouTube videos, which is evidently more relevant than the previous research results conducted by researchers and data analysts. We selected a few well-known videos from “TedX Talks” to make sure the rating system yields accurate results.

Additionally, by restricting the amount of retrieved comments to 100 for every video, each video item's interesting remarks were still highlighted.

Ultimately, we demonstrated that a strong relevance ranking tool for YouTube comments can be constructed using the BERT Transformer and modern State-of-the-Art Pretrained model to make communities more inclusive and solving issues. The input is pre-filtered using a series of pre-processing methods and a classifier in the first step, then topic extraction and a weighted function relevance calculation stage are combined in the next step.

VI. REFRENCES:

[1] Andrei Serbanoiu, Traian Rebedea. ” Relevance-Based Ranking of Video Comments on YouTube.” (2015).

[2] Shubhanshu Shekhar, Akanksha, Aman Saini. ” Utilizing Topic Modelling To Identify Abusive Comments On YouTube.” (2021) International Conference on Intelligent Technologies (CONIT)

[3] Rahul Pradhan. ” Extracting Sentiments from YouTube Comments.” (2021) Sixth International Conference on Image Information Processing (ICIIP)

[4] Mohd Amaan Ansari, Pavan Prajapati, Shivam Dhotre, Sunil Kumar, Sangita Chaudhari. ” Ensemble Learning based Efficient Spam Detection of YouTube Comments.” (2023) 6th International Conference on Advances in Science and Technology (ICAST)

[5] N.M. Samsudin, C.F. Foozy, N. Alias, P. Shamala, N.F. Othman, W.I. Din (2019). "Development of a YouTube Spam Detection Framework using Naïve Bayes and Logistic Regression." Published in the Indonesian Journal of Electrical Engineering and Computer Science. Conference on Advances in Science and Technology (ICAST)

[6] Oshikawa, Ray, Jing Qian, and William Yang Wang. ”A survey on natural language processing for fake news detection.” arXiv preprint arXiv:1811.00770 (2018).

[7] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. ”Latent dirichlet allocation.” Journal of machine Learning research 3, no. Jan (2003): 993-1022.

[8] Dumais, Susan T. ”Latent semantic analysis.” Annu. Rev. Inf. Sci. Technol. 38, no. 1 (2004): 188-230.

[9] Pavlinek, Miha, and Vili Podgorelec. ”Text classification method based on self-training and LDA topic models.” Expert Systems with Applications 80 (2017): 83-93.

[10] Wilbert Wijaya, Made Murwantara, Aditya Rama Mitra. ” A Simplified Method to Identify the Sarcastic Elements of Bahasa Indonesia in Youtube Comments.” (2020) 8th International Conference on Information and Communication Technology (ICoICT)

[11] Pouria Omrani , Zahra Ebrahimian, Ramin Toosi, and Mohammad Ali Akhaee. ” Bilingual COVID-19 Fake News Detection Based on LDA Topic Modeling and BERT Transformer.” (2023) 6th International Conference on Pattern Recognition and Image Analysis (IPRIA)

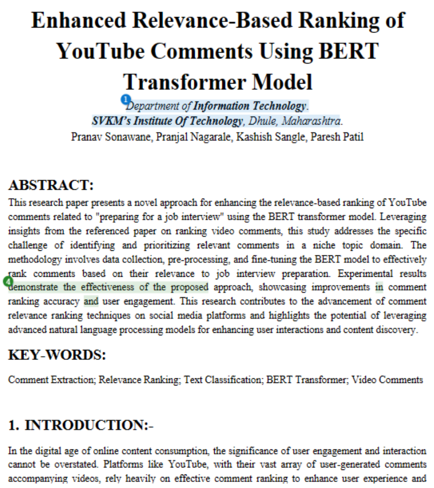
[12] Esteves, Carlos, Christine Allen-Blanchette, Xiaowei Zhou, and Kostas Daniilidis. ”Polar transformer networks.” arXiv preprint arXiv:1709.01889 (2017).

[13] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszko reit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. ”Attention is all you need.” Advances in neural information processing systems 30 (2017).

[14] Adrian M. P., Braşoveanu, Răzvan Andonie. Visualizing Trans formers for NLP: A Brief Survey. IEEE. 2020

[15] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. ”Bert: Pre-training of deep bidirectional transform ers for language understanding.” arXiv preprint arXiv:1810.04805 (2018).

Appendix C: Plagiarism of Research Paper



Enhanced Relevance-Based Ranking of YouTube Comments Using BERT Transformer Model

Department of Information Technology.
SVKM's Institute Of Technology, Dhule, Maharashtra.
Pranav Sonawane, Pranjal Nagarale, Kashish Sangle, Paresh Patil

ABSTRACT:

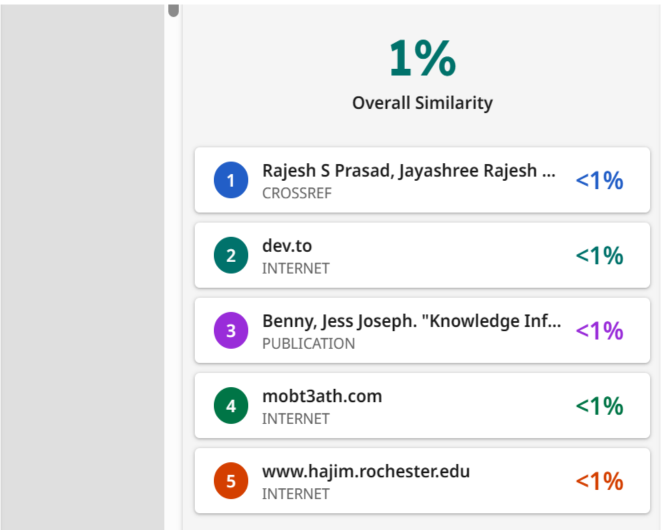
This research paper presents a novel approach for enhancing the relevance-based ranking of YouTube comments related to "preparing for a job interview" using the BERT transformer model. Leveraging insights from the referenced paper on ranking video comments, this study addresses the specific challenge of identifying and prioritizing relevant comments in a niche topic domain. The methodology involves data collection, pre-processing, and fine-tuning the BERT model to effectively rank comments based on their relevance to job interview preparation. Experimental results demonstrate the effectiveness of the proposed approach, showcasing improvements in comment ranking accuracy and user engagement. This research contributes to the advancement of comment relevance ranking techniques on social media platforms and highlights the potential of leveraging advanced natural language processing models for enhancing user interactions and content discovery.

KEY-WORDS:

Comment Extraction; Relevance Ranking; Text Classification; BERT Transformer; Video Comments

1. INTRODUCTION:-

In the digital age of online content consumption, the significance of user engagement and interaction cannot be overstated. Platforms like YouTube, with their vast array of user-generated comments accompanying videos, rely heavily on effective comment ranking to enhance user experience and

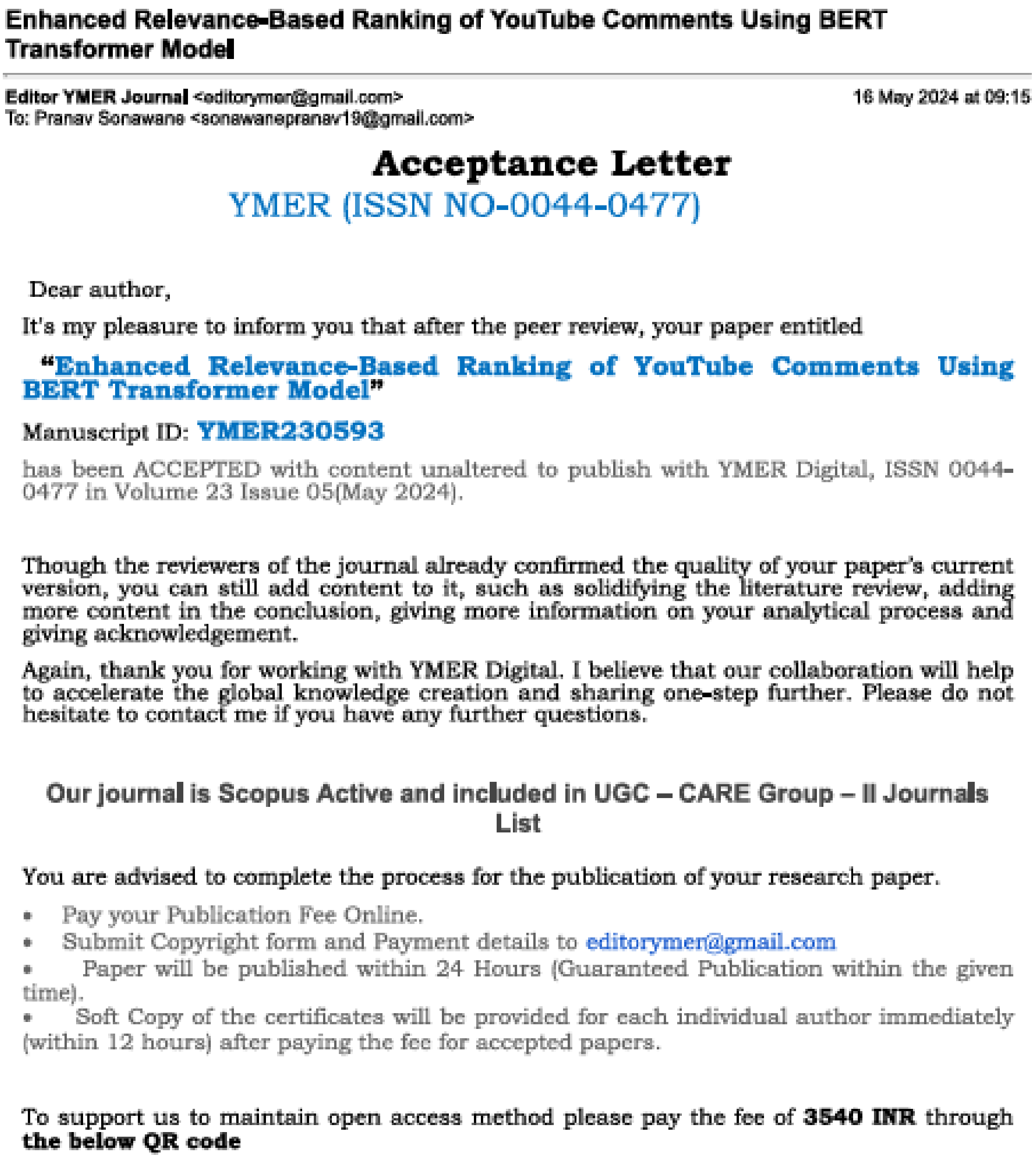


1%

Overall Similarity

1	Rajesh S Prasad, Jayashree Rajesh ...	CROSSREF	<1%
2	dev.to	INTERNET	<1%
3	Benny, Jess Joseph. "Knowledge Inf...	PUBLICATION	<1%
4	mobt3ath.com	INTERNET	<1%
5	www.hajim.rochester.edu	INTERNET	<1%

Appendix D: Acceptance of Published Paper



Enhanced Relevance-Based Ranking of YouTube Comments Using BERT Transformer Model

Editor YMER Journal <editorymer@gmail.com>
To: Pranav Sonawane <sonawanepranav19@gmail.com>

16 May 2024 at 09:15

Acceptance Letter

YMER (ISSN NO-0044-0477)

Dear author,

It's my pleasure to inform you that after the peer review, your paper entitled

"Enhanced Relevance-Based Ranking of YouTube Comments Using BERT Transformer Model"

Manuscript ID: **YMER230593**

has been ACCEPTED with content unaltered to publish with YMER Digital, ISSN 0044-0477 in Volume 23 Issue 05(May 2024).

Though the reviewers of the journal already confirmed the quality of your paper's current version, you can still add content to it, such as solidifying the literature review, adding more content in the conclusion, giving more information on your analytical process and giving acknowledgement.

Again, thank you for working with YMER Digital. I believe that our collaboration will help to accelerate the global knowledge creation and sharing one-step further. Please do not hesitate to contact me if you have any further questions.

Our journal is Scopus Active and included in UGC – CARE Group – II Journals List

You are advised to complete the process for the publication of your research paper.

- Pay your Publication Fee Online.
- Submit Copyright form and Payment details to editorymer@gmail.com
- Paper will be published within 24 Hours (Guaranteed Publication within the given time).
- Soft Copy of the certificates will be provided for each individual author immediately (within 12 hours) after paying the fee for accepted papers.

To support us to maintain open access method please pay the fee of **3540 INR** through **the below QR code**