
Topic Modeling and Sentiment Analysis of YouTube Comments on OpenAI's Stargate Project Using BERTopic and RoBERTa

Yuliana Dewi Proboningrum^{1*}, Hanifah Permatasari¹, Vihi Atina¹

¹University of Duta Bangsa Surakarta, Faculty of Computer Science, Information Systems, Jl. Bhayangkara No.55, Tipes, Serengan, Surakarta, Central Java 57145, Indonesia

Article History

Submitted: 06-05-2026

Revised: 18-06-2026

Accepted: 22-06-2026

Published: 19-09-2026

DOI: 110.32664/j-intech.v14i03.2395

Keywords

BERTopic; Data Center; OpenAI; RoBERTa; Sentiment Analysis; Stargate; Topic Modeling

*Corresponding Author:

220101041@gmail.com

Abstract

The rapid development of generative artificial intelligence has sparked growing interest, widespread adoption across various sectors, and intense public debate, particularly following the announcement of the Stargate project by OpenAI, SoftBank, and Oracle. The diverse public reactions documented on social media platforms, particularly in YouTube comment sections, can serve as a valuable source of public opinion data. The YouTube comment data used in this study was collected through scraping and subsequently processed through a cleaning procedure, yielding clean data covering the period from January 21, 2025, to April 21, 2026. This study aims to identify dominant discussion topics and the underlying sentiment to uncover indications of public concerns and expectations regarding the AI development project led by OpenAI through the Stargate initiative. The analysis in this study employs a structured NLP approach, specifically BERTopic for topic modeling and RoBERTa for sentiment analysis. The RoBERTa model was fine tuned for binary classification using a combination of pseudo labeling, annotation, class balancing, and classification decision threshold optimization, which optimally yielded an accuracy of 0.89 and a macro F1 score of 0.87 in the two class evaluation. BERTopic successfully identified 43 topics with a coherence score of 0.58 and a topic diversity of 0.71.

1. Introduction

Since the emergence of large language models, the development of artificial intelligence has advanced rapidly, leading to the emergence of generative AI capable of automatically generating human-like natural language text, programming code, images, and even videos. These advancements in AI capabilities have had a significant impact across various sectors, ranging from education and industry to finance and healthcare [1]. This has sparked growing interest in and adoption of artificial intelligence, as evidenced by the popularity of generative AI tools such as ChatGPT, Dall-E, Gemini, GitHub Copilot, and similar AI services that are increasingly being integrated into the modern application ecosystem [2]. As the number of users grows both among individuals and within the industry the demand for basic AI computing resources, such as GPUs and data center infrastructure, is also increasing. This trend is further highlighted in a report by Allied Market

Research, which indicates that the AI infrastructure market reached \$23.5 billion in 2021 and is projected to grow to \$309.4 billion by 2031, with a compound annual growth rate of 29.8% [3].

In January 2025, OpenAI, together with SoftBank and Oracle, announced the launch of a large-scale data center development project called Stargate, with an investment of up to \$500 billion [4]. This project aims to meet the massive computing needs of global AI services and strengthen the United States' position in artificial intelligence innovation. Although it is promoted as a driver of economic growth through the creation of hundreds of thousands of jobs and U.S. leadership in the field of AI [4]. This project has sparked a wide range of public reactions, ranging from support to skepticism to outright rejection. Significant public attention to this project is reflected in a Bloomberg Originals YouTube video titled "Inside OpenAI's Stargate Megafactory with Sam Altman" from May 20, 2025, which has been viewed over 6.2 million times one of many YouTube videos covering the Stargate project that clearly captures the attention of millions on this topic. The video features a tour of the construction site, interviews with stakeholders, as well as skeptical and optimistic perspectives from journalists [5], which sparked widespread public discussion in the comments section, reflecting public reactions to this large-scale artificial intelligence project. The public reactions captured in social media comments reflect a range of views, from concerns to hopes [6]. To understand public perception of the project, the use of Natural Language Processing (NLP) is essential. The combined application of topic modeling and sentiment analysis enables the systematic identification of key discussion topics and sentiment trends in public opinion from unstructured and random text, such as YouTube comments [7]. One NLP approach that has consistently demonstrated superior performance is transformer-based methods such as BERT and RoBERTa for sentiment detection, as well as BERTopic for modeling public opinion topics.

The transformer model was first introduced by Vaswani et al. in 2017 in a paper titled "Attention Is All You Need" as a solution to the limitations of earlier models such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTMs) [8][9]. RoBERTa (Robustly Optimized BERT Pretraining Approach) and BERTopic are two methods that utilize the transformer architecture through a self-attention mechanism to understand contextual relationships between words. Previous research has shown that combining transformer-based topic modeling with sentiment analysis can effectively identify public perceptions across various fields. Research by Angelo et al. (2025) [10] shows that the IndoBERT model outperforms traditional models such as Naïve Bayes, SVM, AdaBoost, XGBoost, and LightGBM in sentiment analysis of public discussions regarding the IKN policy across two phases of presidential leadership and transition, when combined with LDA. Another study by Alamsyah & Girawan (2023) [11] also shows that the combination of BERTopic and RoBERTa effectively captures topic structure and sentiment patterns in clothing product reviews, while Ayse Ocal [6] utilizing BERTopic and BERT to explore public perceptions regarding AI development. These findings demonstrate the potential effectiveness of transformer based approaches for extracting meaningful topics and sentiments from large scale text data.

A comparison of these findings provides a solid foundation for further exploration of transformer based methods. Although the specific application of the BERTopic and RoBERTa method combination in uncovering public perceptions regarding large scale AI infrastructure projects such as Stargate remains limited in the existing literature, this represents a research gap that needs to be addressed. Therefore, this study applies BERTopic and RoBERTa to YouTube comment data related to OpenAI's Stargate project to identify dominant discussion topics and uncover indications of public concerns and expectations regarding the latest developments in AI infrastructure.

2. Research Method

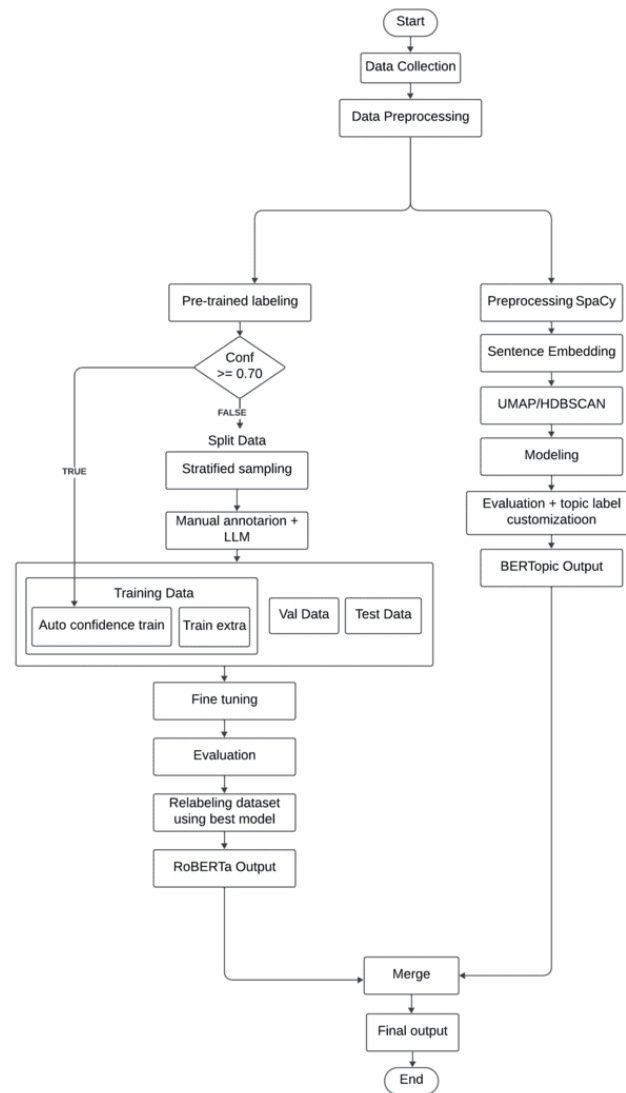


Figure 1. Research Method

This study employs a two method approach that combines topic modeling and sentiment analysis to systematically examine public discussions in YouTube comment sections regarding OpenAI’s Stargate project. The overall workflow is shown in Figure 1 above and is broken down into five sequential stages: data collection, data preprocessing, parallel modeling (BERTopic and RoBERTa modeling), integration of the results from both analyses, and interpretation of the findings, as described in the following methodology subsection.

2.1 Data Collection

The data was obtained and collected from numerous YouTube videos using the YouTube Data API v3, targeting comments on videos that included various keywords relevant to the research topic, such as “Stargate Data Center,” “OpenAI Stargate Project,” “Stargate AI Infrastructure,” “Stargate \$500B,” “Stargate AI News,” and “OpenAI Data Center Stargate.” The raw dataset consists of 297,740 entries from 80 unique videos before undergoing any filtering or further processing. The data comprises the following columns: video_id, published_at, video_title, channel, keyword, and comment. YouTube comments were selected due to their ease of access and the abundance of data rich in public opinion.

2.2 Data Preprocessing

The raw comment dataset undergoes a cleaning process to ensure data quality before analysis. The data transformations applied during preprocessing include:

- 1.) Video Selection
The raw data obtained from the scraping process underwent a cleaning process to remove comments on irrelevant videos, the filtering process for irrelevant videos resulted in a clean dataset of 58 relevant video titles covering the period from January 21, 2025, to April 21, 2026.
- 2.) Removal of Irrelevant Data
The removal of irrelevant data includes the deletion of emojis, empty values, duplicates, links/URLs, mentions, symbols, punctuation marks, repeated characters, as well as numbers and timestamps.
- 3.) Casefolding
The process of converting all letters in a text to a uniform form, namely lowercase. The goal is to reduce confusion between words that actually have the same meaning, simply because of variations in the letters used [7].
- 4.) Normalization
The normalization stage involves steps to convert text into a standard, uniform, and consistent form, as well as converting non-standard text into a standard form so that it can be widely understood [7]. The normalization stage in this study includes the normalization of slang and the development of contractions.

Once the cleaning process is complete, each data row is assigned a unique identifier in the form of an additional column, `id\_data`, which serves as a permanent key for combining the results of sentiment analysis and topic modeling in the next stage. The total number of data points after the preprocessing stage is 54.403 comments, which will then be used as combined input for the sentiment analysis and topic modeling workflows running in parallel.

2.3 RoBERTa Pipeline

2.3.1 Labeling Sentiment with Pre-Trained Model

Initially, none of the collected comments had sentiment labels, so we used a pre-trained model for labeling. The pre trained model used in the pseudo-labeling step was cardiffnlp/twitter-roberta-base-sentiment, a RoBERTa model specifically trained on Twitter social media data that reflects the characteristics of informal, concise language, slang, emojis, and the typical noise found in social media. [12], so that it closely matches the characteristics of YouTube comments. This pseudo labeling approach is used to generate initial labels, thereby reducing the annotation burden without completely sacrificing label quality [13]. After the pseudo-labeling process, the data is then separated based on confidence scores. Comments with scores of ≥ 0.70 are stored in the auto confidence pool, while those with lower scores are directed to the stratified sampling stage, this is because including data that the model is unsure how to label in the training subset increases the risk of the model learning incorrect labels. The table below shows the distribution of the number of data points per class along with their confidence scores.

Table 1. Confidence Score Distribution

Class	Total	Confidence < 0.70	Percentage
Negative	28.738	9.349	32,5%
Neutral	18.303	12.563	68.6%
Positive	7.362	2.613	35.5%

Based on the table, it is clear that the neutral class dominates the low confidence scores, indicating that the model has greater difficulty distinguishing neutral sentiment compared to other classes.

2.3.2 Split Data

From the group of comments with low confidence scores, stratified random sampling was performed to create an annotated subset, ensuring equal representation across all categories (negative, neutral, positive). This technique ensures that data is drawn proportionally from each class to produce a balanced

representation [14]. The number of stratified sampling draws that need to be annotated from each subset is presented in Table 2 below.

Table 2. Stratified Sampling Result

Class	Set	Total
Negative	Train	1.600
Neutral		1.600
Positive		1.600
Negative	Val	266
Neutral		267
Positive		267
Negative	Test	134
Neutral		133
Positive		133

The selected samples were then annotated using a GPT based LLM, which generated pseudo sentiment labels for all samples in the low confidence stratum. The prompt structure used to guide the LLM’s labeling was adapted from a prompt framework developed in previous research [15]. The prompt provides a task instruction to “Classify” by defining which emotions each class must contain for example, the positive class must include emotions or tones of support, appreciation, and optimism, while the negative class must include criticism, anger, rejection, or other negative emotions, and the neutral class must include comments that are neither positive nor negative in tone, such as those conveying information or facts [15]. The use of these standardized prompts aims to maintain consistency in the labeling process. To assess the reliability of the LLM annotations, a random sample of 150 items was drawn from the annotated dataset and then annotated independently by a single person without referring to or viewing the labels generated by the LLM, in an effort to avoid annotation bias. The inter rater agreement between the GPT based annotations and the author’s annotations was calculated using Cohen’s kappa, yielding a value of 0.62, which, based on research by [16] a Kappa value in the range of 0.61 to 0.80 falls within the substantial range, indicating that the data has been labeled consistently and reliably.

Once the annotation process is complete, the neutral labeled data is temporarily separated from the rest of the dataset to preserve the integrity of the binary classification during the fine tuning process, as including all three classes including the neutral class could reduce accuracy due to the high ambiguity in the neutral class.

Table 3. Class Distribution after Annotation

Class	Set	Total
Negative	Train	2.509
Positive		961
Negative	Val	407
Positive		150
Negative	Test	198
Positive		98

2.3.3 Fine Tuning RoBERTa

1) Class Balancing

Prior to the fine tuning process, the binary training dataset exhibited significant class imbalance, with negative instances accounting for 72.3% and positive instances only 27.7%. Two strategies were applied sequentially to address this issue. First, the training set was augmented with pseudo labeled samples from the high confidence group (confidence ≥ 0.90), with a maximum sampling limit of twice

the size of annotated sample specifically, twice the number of pseudo labels ≥ 0.90 from the negative training class, totaling 7.527, and twice the number of pseudo label ≥ 0.90 from the positive training class, totaling 2.883. Second, an undersampling technique is applied to the majority class (negative) to equalized the number of minority class, resulting in a balanced final training dataset [17]. The distribution of high confidence scores in the pseudo labeling group and the sentiment distribution from manual annotation are presented below.

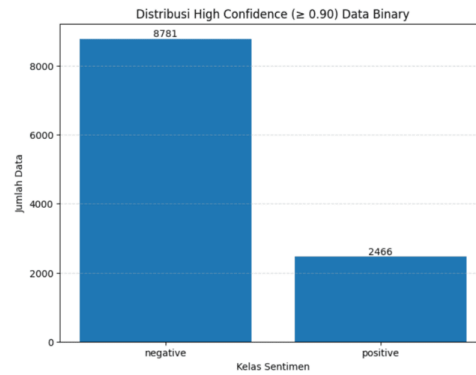


Figure 2. High Confidence Score Pseudo Label Distribution

Table 4. Data Balancing with Undersampling

Sentiment	Before Data Balancing	After Data Balancing
Negative	2.509	2.883
Positive	961	2.883

The combination of augmentation and undersampling has been shown to produce more stable classification performance on imbalanced datasets than either strategy applied singly [17] [11].

2) Model Architecture and Training Configuration

The sentiment analysis was performed using RoBERTa, an optimized BERT variant pre trained on a large corpus of English, including English Wikipedia, CC-News, OpenWebText, and Stories, yielding better contextual word understanding than BERT [11], In this step, all texts were tokenized with a maximum length of 256 tokens and processed in batches of 16. The model was fine tuned using 10 epochs with the following configuration:

- Learning rate: 5×10^{-6}
- Label smoothing: 0.1 to reduce model overconfidence by preventing the target class from being overconfident [18].
- Early stopping: A patience count of 3 epochs was monitored for validation loss, with the optimal checkpoint kept at the epoch that achieved the minimum validation loss.

3) Optimization Decision Threshold for Classification

Model optimization does not rely on the default classification decision threshold of 0.50, but instead systematically searches for a threshold among candidate values ranging from 0.30 to 0.70, using the macro score as the optimization criterion. The search for the optimal classification decision threshold is performed on the validation data to prevent data leakage if it were performed on the training data. The results of the search for the classification decision threshold are presented in the following table.

Table 5. Optimal Decision Threshold

Threshold	F1 Negative	F1 Positive	Macro F1
-----------	-------------	-------------	----------

0.30	0.8708	0.7328	0.8018
0.35	0.8816	0.7458	0.8137
0.40	0.8807	0.7407	0.8107
0.45	0.8837	0.7450	0.8143
0.50	0.8869	0.7478	0.8173
0.55	0.8915	0.7529	0.8222
0.60	0.8889	0.7372	0.8130
0.65	0.8937	0.7407	0.8172
0.70	0.8950	0.7325	0.8137

Based on the process of finding the optimal threshold, it can be observed that the default threshold (0.50) yields a negative F1 score of 0.88, which is already above average, while the positive F1 score is only 0.74, indicating that the default threshold is not yet optimal. Increasing the threshold from 0.50 to 0.55 improves the positive F1 score from 0.74 to 0.75, with the macro F1 score also reaching its highest value. Thresholds ranging from 0.65 to 0.70 consistently yield a further increase in the negative F1 score but instead lower the positive F1 and macro F1 scores; thus, it can be concluded that the optimal threshold is 0.55. This threshold search technique is based on previous research that examined methods for optimizing model performance with imbalanced data without relying solely on the default threshold, which found that the Threshold Decision Calibration method outperforms SMOTE and Class Weight [19] [20].

2.3.4 Evaluation RoBERTa

1) Precision

$$Precision = \frac{TP}{TP+FP} \quad (1)$$

2) Recall

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

Description:

TP/True Positive : Positive class data classified as positive

TN/True Negative : Negative class data classified as negative

FP/False Positive : Negative class data classified as positive

FN/False Negative : Positive class data classified as negative

3) F1-Score

$$F1 = 2 \frac{Precision \times Recall}{Precision + Recall} \quad (3)$$

The fine tuned RoBERTa model is evaluated on the test set using a confusion matrix and the evaluation is performed using the optimal threshold.

2.3.5 Relabeling Dataset using Best Model

The model with the best performance was then used to label the entire dataset, which comprised 54.403 comments. To generate a three-class output, a confidence threshold of 0.75 was applied. Comments with a high predicted class probability that met or exceeded this threshold were assigned a binary prediction label (negative or positive), while comments below the threshold were categorized as neutral because the model lacked sufficient confidence to classify them as either negative or positive. This mechanism is based on the principle of selective prediction in NLP, which states that a classification system may withhold or adjust its prediction when the model's confidence estimate falls below the threshold to ensure the reliability of results on the final dataset [21]. The study shows that the level of prediction confidence is positively correlated with model accuracy, so predictions with low confidence tend to represent ambiguous data.

In this study, the neutral class does not originate from the model training process because the model is trained and evaluated using a binary classification scheme (negative and positive). Instead, the neutral class serves as an indicator of the model's uncertainty regarding the two learned classes. Empirically, comments

with low confidence scores are generally informative statements, questions, or comments that do not indicate a strong sentiment bias. This approach is a deliberate trade-off to maintain the quality of negative and positive labels while avoiding the assignment of less-certain sentiment labels during the inference stage. The implications of this approach for interpreting sentiment distributions are discussed further in subsection 3.2.4.

2.4 BERTopic Pipeline

2.4.1 SpaCy Preprocessing

Before topic modeling was performed, the cleaned comments underwent an additional preprocessing stage using the SpaCy NLP library. Each document was processed through lemmatization and filtering based on parts-of-speech (POS), retaining only tokens with high semantic relevance, such as nouns, verbs, and adjectives, which can improve topic interpretability, produce more specific topics, and reduce noise in BERTopic results [22] [23]. In this study, topic modeling only used nouns, propn, and adjectives. Documents that produced fewer than three words after filtering were excluded from topic modeling, resulting in a final dataset of 45,724 documents. The `id_data` identifier was retained throughout this process to allow merging with sentiment results in a later stage.

2.4.2 Sentence Embedding

Before modeling topics using BERTopic, the data first undergoes a sentence embedding process, in which documents are encoded into vector representations using the `all-mpnet-base-v2` sentence transformer model. This model was chosen because of its superior performance compared to the MiniLM model.

2.4.3 UMAP & HDBSCAN

Once the documents have been converted into vector format that the model can understand, the next step is to perform two sequential operations prior to the modeling stage, configuring UMAP (Uniform Manifold Approximation and Projection) and HDBSCAN (Hierarchical Density Based Spatial Clustering of Applications with Noise). UMAP is applied to reduce the embedding dimension to a lower dimension, while HDBSCAN is used to identify/discover coherent and context based topic clusters [22] and to model low density or dissimilar regions as noise topics (topic -1). In this study, UMAP was applied with the configuration `n_neighbors=10`, `n_components=5`, `min_dist=0.1`, and the cosine metric. Furthermore, HDBSCAN was applied with the configuration `min_cluster_size=90` and `min_samples=10`.

2.4.4 BERTopic Modeling

At this stage, modeling was performed using BERTopic, a neural topic modeling framework that integrates contextual embedding, density based clustering, and c-TF-IDF-based term weighting to generate semantically coherent topics [19]. The application of BERTopic in this study is configured with `CountVectorizer` parameters set to `min_df=5`, `max_df=0.80`, and `ngram_range=(1, 2)`, enabling the detection of informative bigrams such as “data center” or “artificial intelligence.” Furthermore, the number of topics is set to “auto” to allow the model to organically determine the optimal number of topics from the data structure. Subsequently, to refine topic representation and simplify the process, three representation models were used on top of the c-TF-IDF scores: KeyBERTInspired for semantic keyword ranking, Maximal Marginal Relevance (MMR, `diversity=0.8`) to ensure keyword diversity and minimize redundancy within each topic, and Part-of-Speech filtering as the final layer to enforce the grammatical relevance of representative keywords [24].

2.4.5 Evaluation and Topic Label Customization

The quality of the BERTopic model was evaluated using two metrics, a coherence score, which calculates the semantic similarity between keywords, and topic diversity, which calculates the ratio of unique words by measuring the proportion of unique words that rank at the top across all topics [25]. After model evaluation, the topic labels generated by BERTopic were replaced with descriptive labels to make them easier to

understand. The labeling process was carried out with the assistance of a GPT based LLM [26], this mechanism is based on research findings indicating that labeling topics with interpretations in the form of sentences improves readability and ease of interpretation compared to labels based on raw keywords [24][26]. The topic labeling process is carried out by simultaneously integrating the three keyword extraction results from BERTopic namely, KeyBERT, MMR, and POS so that the resulting labels improve interpretability [24].

2.4.6 BERTopic Output

The topic modeling pipeline produces a structured output dataset for each of the 45.724 valid documents, including topic IDs, topic classifications, custom description labels, and topic cluster membership probabilities. This dataset is keyed using `id_data` to allow integration with sentiment analysis results.

2.4.7 Data Merging

The output of the parallel flow analysis, RoBERTa sentiment labels results (54.403 comments) and BERTopic analysis results (45.724 documents) were integrated through an inner join operation using `id_data` as the connecting key [6]. All 45.724 documents in the BERTopic output were successfully matched with their corresponding sentiment labels in the RoBERTa output. Outlier documents were then excluded from the merge analysis, resulting in a final analysis information set of 45.535 documents each containing both topic labels and sentiment labels.

2.4.8 Findings/Final Output

The integrated dataset of the two models served as the basis for a combined sentiment topic analysis. For each generated topic, the sentiment distribution was calculated as the percentage of negative, positive, and neutral documents. For each topic, the five most representative documents were extracted based on the highest RoBERTa confidence score within the dominant sentiment class to allow for qualitative verification of topic coherence and sentiment accuracy for further interpretation.

3. Result and Discussions

3.1 RoBERTa Sentiment Analysis

The RoBERTa model was evaluated using an optimized decision threshold of 0.55, obtained through a decision threshold search technique designed to improve accuracy. This technique was chosen because it yields the most balanced and optimal F1 score for binary classification results (positive and negative). The purpose of this threshold search technique is to refine the model's performance evaluation results. The results confirm that a threshold of 0.55 is the optimal point for the model to classify the positive class more strictly. The model was evaluated using a test dataset with the optimal threshold set at 0.55, and its performance can be seen in Figure 3 below.

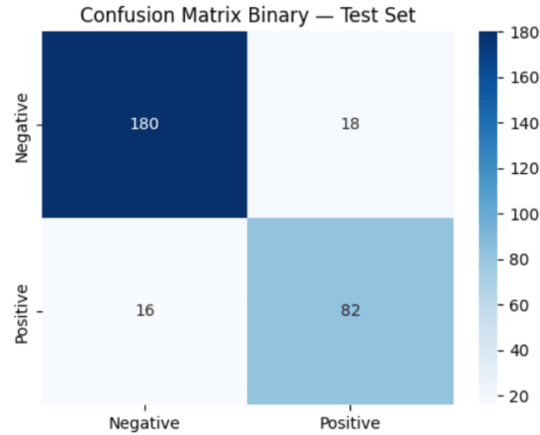


Figure 3. Confusion Matrix

Based on the confusion matrix, the model successfully classified 180 negative sentiment data points correctly (true negatives) and 82 positive sentiment data points correctly (true positives). Meanwhile, 18 negative sentiment data points were incorrectly predicted as positive (false positives) and 16 positive sentiment data points were incorrectly predicted as negative (false negatives). The error distribution shows a balanced trend between false negatives and false positives, indicating that the model does not lean too heavily toward any particular class.

	precision	recall	f1-score	support
Negative	0.92	0.91	0.91	198
Positive	0.82	0.84	0.83	98
accuracy			0.89	296
macro avg	0.87	0.87	0.87	296
weighted avg	0.89	0.89	0.89	296

Figure 4. Evaluation Model Score

The evaluation results show that the process of refining the RoBERTa model yields excellent classification performance. Although the dataset exhibits significant sample imbalance, the model successfully demonstrates resilience against majority class bias. This is evidenced by a macro F1-score of 0.87, which is close to the accuracy score (0.89). Meanwhile, the minority class (positive) still achieved an F1 score of 0.83, indicating that the model is capable of recognizing positive patterns well without being overly dominated by negative samples, suggesting that the data balancing process in the previous stage was highly effective.

3.2 BERTopic Topic Modeling

The evaluation of the BERTopic model yielded 43 topics with a coherence score (C_v) of 0.58 and a topic diversity score of 0.714, indicating a low level of overlap among the generated topics [25]. This coherence score is generally classified as moderate, indicating that the keywords within each topic exhibit a good level of semantic relevance. To assess the validity of this score, the results were compared with several previous studies that also utilized BERTopic. Alamsyah & Girawan [11] achieved a higher coherence score of 0.67 in product reviews that used more consistent vocabulary around a more focused topic related to clothing, while Samsir et al. [27] reported coherence scores ranging from 0.42 to 0.58 (average 0.51) for academic abstract data, indicating that this range reflects strong semantic consistency within each identified topic. The score of 0.58 obtained in this study is lower than the value achieved in the study by [11] [25] due to differences in the characteristics of the datasets used, yet it falls within the upper limit of the range reported in studies published by [27] and by [7] which examined YouTube comments regarding the Free Nutritious Meals policy.

3.2.1 Intertopic Distance Map

Intertopic Distance Map is a graph that shows how close or far the semantic similarity is between topics [7]. Figure 5 below presents an intertopic map containing the distance between the generated topics. Topic 0 dominates with the largest circle size, indicating the widest document coverage. There is a cluster of adjacent and overlapping topics in the upper left quadrant filled by topic clusters 0, 1, 6, 22, 24, 27, 38, 42 and clusters 4, 8, 13, 16, 30, 33, 34, indicating semantic closeness between these topics. Meanwhile, several other topics appear isolated in the distant quadrant, indicating that these topics have more unique and different vocabulary and context than most of the other topics.



Figure 5. Intertopic Distance Map

3.2.2 Top Word Scores

The Top Word Scores visualization in Figure 5 below shows the intensity of keywords that best represent a topic, based on their c-TF-IDF scores. The higher the score, the more frequently the word appears and represents the topic.

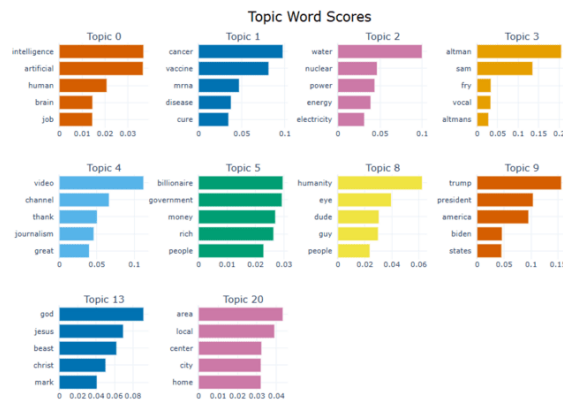


Figure 6. Top Word Scores

An example of top word score interpretation is topic 0, which is dominated by the words "intelligence" and "artificial," with the highest scores among all keywords, followed by "human," "brain," and "job." This pattern confirms that discussion of this topic centers on the relationship between artificial intelligence and human thinking capacity, including the dominance of artificial intelligence in the job market.

3.2.3 Topic Results

Topic results are the section that shows the topics generated from the BERTopic analysis. Based on the BERTopic parameter configuration used, BERTopic generated 43 topics. The table below presents all the topics generated from the topic modeling analysis applied to a large dataset of public comments regarding the OpenAI Stargate project, showing topic labels, the distribution of the number of topics identified in the document collection, keywords (words representing the topics), and representative documents.

Table 6. Topic Modeling Results

No	Topic	Distribution	Keywords	Representative Docs
0	Artificial Capabilities and Impact	11.949	intelligence, artificial, human, brain, job, people, thing, computer cancer, vaccine,	at this rate ais will soon reach sentients within decades
1	mRNA Vaccines Cancer Discourse	1.800	mrna, disease, cure, health, medical, healthcare	they should be banned period they endanger the publics health does not create too many jobs
2	Water Energy Nexus in Data Center	2.281	water, nuclear, power, energy, electricity, center, heat, plant	also cover the serious damage it is doing to every human that lives in its vicinity making water undrinkable and other hazards
3	Sam Altman Public Perception Controversy	1.521	altman, sam, fry, vocal, altmans, voice, scam, guy video, channel,	i do not trust sam altman something is just off with this guy
4	YouTube Investigative Journalism Channel Content	1.564	thank, journalism, great, content, youtube, fake billionaire,	truly excellent piece bloomberg originals thank you for consistently putting out quality stories
5	Oligarchic Capitalism and Wealth Inequality	3.120	government, money, rich, people, world, wealth, country openai, google,	we should never have allowed this in our country it is a poison to our planet
6	OpenAI Governance and Commercialization	978	microsoft, gemini, open, response, company, model voice, vocal, fry,	we could pull the plug on it right now it is just those few bad people that will keep it going maybe openai should be shutdown
7	Vocal Fry Speech Pattern Criticism	934	loud, annoying, speech, guy, audio	anyone who speak in a way as if they have spikes in their throat are a red flag
8	Perceived Threats to Humanity	1.845	humanity, eye, dude, guy, people, scary, shit, human	needs to shave his head it is scaring me more his mannerism is strange almost moves his head like a robot
9	Public Opinion on Presidential Leadership	1.192	trump, president, america, biden,	like everything else man made has and will destroy peoples lives the out come is always the same wtg america

No	Topic	Distribution	Keywords	Representative Docs
10	Elon Musk and Tech Billionaires Scrutiny	920	states, united, donald, country elon, musk, thiel, peter, altman, sam, zuckerberg, trump	what the fuck happened to his legs they look like magnetos
11	Stargate Program and Remote Viewing	806	stargate, project, remote, program, cia, star, psychic, gate	i wonder how much control of this vests with the deep state this highly dangerous project is being sold to the public under its guise of improving public health but that is just the tip of the iceberg of what can and probably will sink humanity itself o
12	Data Center Infrastructure and Development	945	data, center, datacenter, centre, datum, power, cloud, area	digging for ground is dangerous once the water goes dry the land will die and more earthquakes means data centers explode and lost everything or it became desert united states of america is klling it self for money
13	Biblical Prophecy and End Times Theology	1.163	god, jesus, beast, christ, mark, bible, revelation, spirit	we do not want it i never consented to having my body and mind hacked hands off of the little ones you have poisoned us all to hack us to mind control us just stop it and let us be let us live our yrs we are dieing quick enough stop poisoning us we cannot breathe we cannot drink god water we cannot...
14	Evaluation of Interview Quality and Performance	752	interview, interviewer, question, sam, chris, great, anderson, answer	these people she is interviewing are crazy people who are going to make us slaves and they honestly look like they know thta behind their eyes which are also kind off soulless was it just me that felt like mr hardhat was staring on her forehead to the interviewers credit she was pushing a little bit...
15	ChatGPT Usage and User Interaction	780	chatgpt, chat, gpt, user, version, free, question, time	it is utter hot garbage and i cancelled my account because of it also is he using chatgpt to answer these questions what the everloving heck how do you not answer a question and yet talk for so long
16	Egg Prices and Food Economics	407	egg, basket, video, dozen, price, expensive, waste, food sam, question,	your eggs are a lie they are all plastic
17	Interpersonal Perceptions of Sam	688	sister, chris, guy, good, conversation, host	sam feel like chatgpt whie you ask it to how make nitroglycerin it just do not give you the answer
18	Geographic Distribution of Data	925	datum, center, power, energy,	okay those construction costs for the empty data centres are way too high

No	Topic	Distribution	Keywords	Representative Docs
	Centers		centre, water, data, people	even if they include off site infrastructure costs
19	Evaluation of GPT Model Performance	402	gpt, model, well, question, claude, version, user, thing	human body has a unique way of handling unused organs and body parts its just reduces the flow of blood and essential nutrients there chat gpt will be a major because of dummer people
20	Data Center Development and Local Impacts	1.022	area, local, center, city, home, land, town, tax	in years there will no pilots or air traffic controllers it will just be too risky
21	Environmental Pollutions from Data Centers	577	noise, sound, pollution, vibration, wall, center, frequency	its slowly poisoning the population and messing up our natural frequency its one of the reasons everything feels off on top of that those weird towers i am seeing pop up everywhere needs to be taken down
22	GPU Market and Hardware Economics	400	gpu, nvidia, chip, ram, price, card, hardware, graphic technology, bro,	the people that are obsessed with cpus are the reason we are at where we are at
23	Tech Elites and Technological Disruption	924	tech, technological, revolution, industrial, people, machine artist, art, creative, copyright, creativity, style, work, theft	not only will the tech this guy relies on destroy the worlds ecosystems but it will literally make everyone stupid it is a twofold enshittification of the planet
24	AI and Creative Intellectual Property	504		soon as it becomes a real threat governments will pull the plug just keep it for military use
25	China in Global Geopolitical Competition	478	china, united, states, chinese, america, russia, country, deepseek	this is not the way we should be living the deep state is already attacking people with deus what do you think au will do to humanity i do not agree with becoming another china to be better and faster than china why are these criminal corporations being given so much power to destroy this planet the...
26	Generational Perspectives on Digital Technology	886	computer, internet, phone, people, young, obsolete, old, year	that is not a firearm it is tool if any gun made pre or it is classified as not a firearm
27	AGI Development Feasibility and Implications	438	agi, llm, asi, human, current, model, intelligence, possible	did he just say to get agi to be built here timeline agi is crazy scary and now they have b to construct it wow beam me up scotty i want off the planet
28	Ponzi Scheme and Financial Fraud	433	scam, scheme, ponzi, history, fraud, scammer, money, big	history has proved many many times never to trust this tribe

No	Topic	Distribution	Keywords	Representative Docs
29	Public Perception of Tech Executives	228	larry, ellison, oracle, old, trump, citizen, year, behavior	mr larry is talking about injecting people with something that will kill there awareness there righteousnees what he calls cancer is really the god in us all there so creepy fuck them
30	Reflections on Future Time Horizons	624	year, long, decade, happy, future, ready, people, comment	shameful these companies should be heavily fined or shut down hell we are all going to die so what is the point it just show the companies with money can do whatever they want
31	Fictional AI Apocalypse Narrative	250	skynet, terminator, connor, matrix, john, sky, movie, sarah cleo, interview, question, great,	we are putting too much stock into artificial intelligence we ca t let it get too big ever heard of skynet artificial intelligence is scary as hell to me anyone else
32	Audience Receptions of Interview Content	280	amazing, answer, sam, conversation	cleo your questions were so incredibly well articulated sam was super engaged and challenged making for the most compelling interview
33	Perceived Time Wastage in Content	557	minute, ass, time, hour, waste, word, late, guy	it is about time to start blowing these centers up all this requests are a waste of time
34	Epistemology of Truth and Evidence	712	answer, truth, fact, question, hard, people, good, real	where are the books and movies like that tell us the truth to make some aware it is real and its is coming it will be worse than we can imagine because stupid mistakes gets rushed
35	Dystopian Science Fiction Cinema	411	movie, dystopian, scifi, fiction, science, report, gattaca, gear	think about it i atlas could be a garbage dump a dangerous mass of highly toxic and contaminating metals not a vehicle but a mass of radiant and biologically hazardous material
36	Economic Bubbles and Market Crashes	339	bubble, burst, crash, stock, boy, market, economy, company	it is no longer a bubble it is a huge infected wound that is slowly killing us
37	Israeli Palestinian Political Discourse	263	jew, israel, jewish, jews, gay, gaza, zionist, genocide	they all seem like actors something is not right with this video
38	Quantum Computing Power	202	quantum, computing, computer, power, consciousness, artificial, intelligence, synthetic	this right here is humanities downfall it is literally pure evil and when you combine it with quantum computing it is almost inevitable to lead to global conflict of incredible magnitude and likely the destruction of our species humans highly likely instigated by the artificial intelligence tha...
39	Climate Change and Environmental	403	climate, change, global,	the ground will collapse in on itself if the water is sucked up out the ground this is

No	Topic	Distribution	Keywords	Representative Docs
	Impact		warming, carbon, footprint, energy, green	crazy they are just destroying the climate it is disgusting
40	Whistleblower Case and Tech Controversy	194	suchir, balaji, yang, suicide, murder, employee, justice	why cannot we just fix the world as it is accelerationism is suicide without the masses consent once again
41	Terminator and AI Apocalypse Cinema	195	terminator, movie, matrix, human, series, robot, intelligence, cop	when is terminators human society is destroying itself
42	Eric Schmidt on AI and Labor	243	schmidt, eric, programmer, job, year, schmidts, human, agent	we are being sucked into extinction mode and we are absolutely clueless

Based on the modeling results, the 43 topics identified reflect the diversity of public discourse regarding OpenAI’s Stargate project. The “Artificial Capabilities and Impact” topic had the largest document volume (11.949), dominated by keywords such as “intelligence,” “artificial,” “human,” “brain,” and “job,” indicating discussions centered on the relationship between artificial intelligence and human capabilities, including its impact on the labor market. Topic 5, “Oligarchic Capitalism and Wealth Inequality,” is the second largest topic (3.120 documents), reflecting public perceptions regarding the concentration of economic power in the hands of a select few due to technological acceleration. Topic 2, “Water Energy Nexus in Data Center,” as the third largest topic (2.281 documents), highlights massive energy and water consumption and resource sustainability issues resulting from data center construction.

The impact of data center infrastructure on the environment and local communities is reflected in Topic 12, “Data Center Infrastructure and Development” (945 documents); Topic 20 “Data Center Development and Local Impacts” (1.022), Topic 21 “Environmental Pollutions from Data Center” (577 documents); and Topic 39, “Climate Change and Environmental Impact” (403 documents). These topics consistently highlight the impacts of pollution, disruptions to communities, and contributions to climate change. On the other hand, several topics specifically discuss direct experiences with AI products and the quality of information delivery, such as Topic 4: YouTube Investigative Journalism Channel Content, Topic 14: Evaluation of Interview Quality and Performance, Topic 15: ChatGPT Usage and User Interaction, Topic 19: Evaluation of GPT Model Performance, and Topic 32: Audience Receptions of Interview Content. These topics differ in nature from the previous ones because they focus more on evaluating the direct use of AI products and the content received.

A number of other topics reflect how public discussions about Stargate intersect with broader issues, such as topics 9 and 25, which touch on the domestic and global political dimensions of technological competition; topics 28 and 36, which highlight financial concerns; and topics 11, 30, 34, 38, and 40, which reflect skepticism, conspiracy theories, and distrust regarding the information shared about this large-scale project. Several other topics, such as topics 1, 11, 13, 16, and 37, appear to have no direct connection to the OpenAI Stargate project. These topics were retained in the modeling results because BERTopic is an unsupervised learning model that maps topics purely based on semantic density through HDBSCAN and the c-TF-IDF procedure, so the resulting topics reflect the structure of the data as it is [24]. Topic 11 highlights the conspiracy theory regarding the similarity in naming between OpenAI’s “Stargate” data center project and other projects. Meanwhile, in Topic 16, the majority of comments used the figurative meaning of the term

“eggs,” which is closely associated with food economics such as the saying “don’t put all your eggs in one basket” as a metaphor to criticize the risk of concentrating massive investments, specifically the \$500 billion allocated to OpenAI’s Stargate project.

3.2.4 Sentiment Distribution on Topics

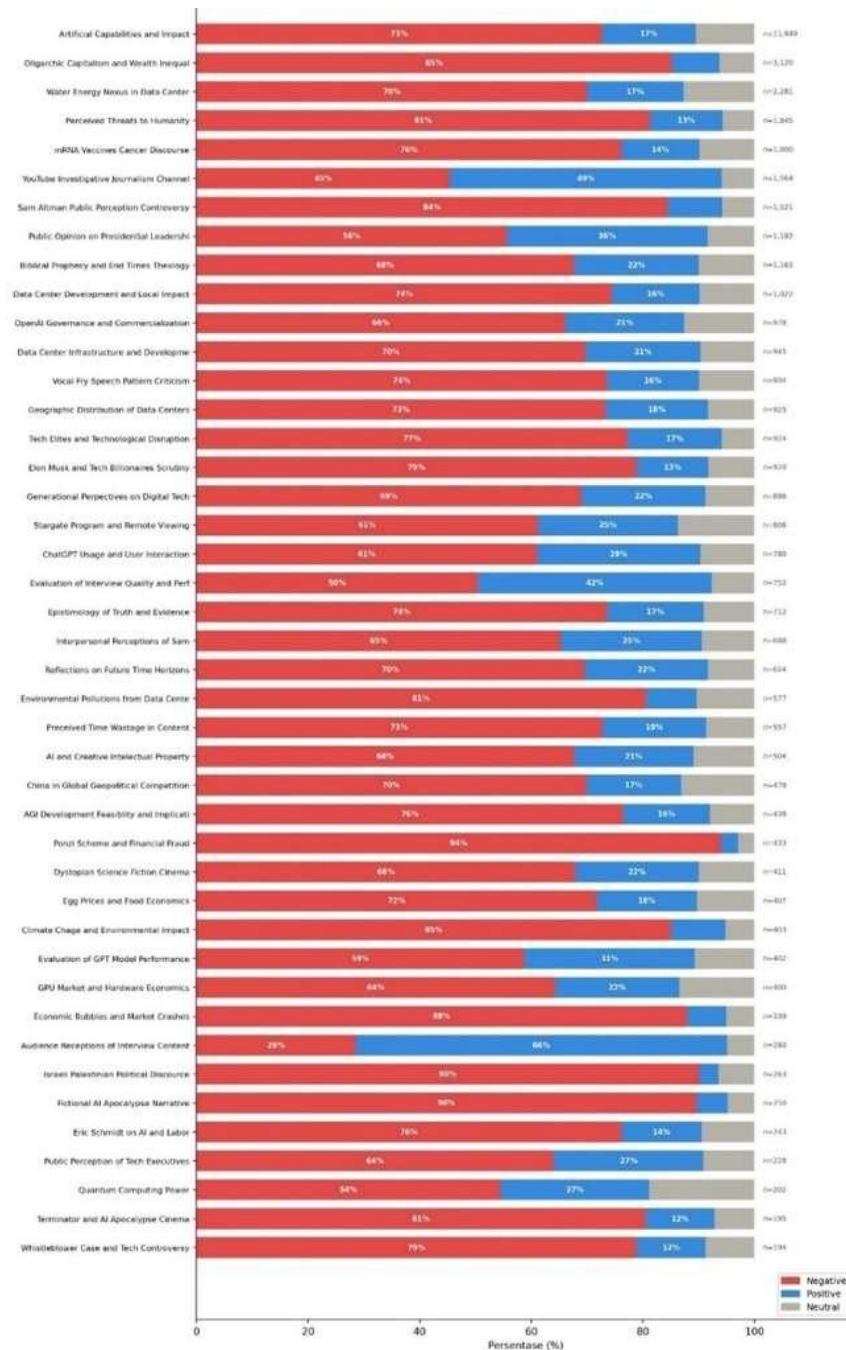


Figure 7. Sentiment on Topic Distributions

Before interpreting the sentiment distribution in Figure 7, it is important to understand that the neutral class in the final labeling results does not represent a true neutral sentiment prediction from the training process, but rather a signal of model uncertainty resulting from the trade off described in subsection 2.3.5 to maintain

the quality of the negative and positive classes, where the model was trained and evaluated solely on a binary classification scheme, whereas the neutral class in the final output purely represents the model's confidence level falling below the aforementioned threshold. Thus, the neutral proportion for each topic in Figure 7 is interpreted as the level of ambiguity in the comments regarding that topic. For example, the topic "Quantum Computing Power," which has a neutral proportion of 19%, reflects comments that merely convey information without leaning toward either negative or positive sentiment, such as the example comment "stargate is just another version of the hadron collider at CERN but with super advanced artificial intelligence and quantum computing capabilities" on the "Quantum Computing Power" topic.

The overall visualization in Figure 7 has addressed the research objective, which is to identify topics of public discussion related to the development of OpenAI's Stargate project and their sentiment trends. Of the 43 topics identified, the majority were dominated by negative sentiment. Some of the topics with the highest levels of negative sentiment are topics 28 (94%), 37 (90%), 31 (90%), 39 (85%), 5 (85%), and 36 (88%). These results indicate that the negative sentiment driving public concern regarding the Stargate project is not only related to AI technology itself but also links it to economic, political, and environmental issues, as well as conspiracy theories circulating in society.

Public concerns are most evident in three main topic groups. First, the group of topics related to technological threats and the future of humanity such as "Artificial Capabilities and Impact," "Perceived Threats to Humanity," "AGI Development Feasibility and Implications," and "Terminator and AI Apocalypse Cinema" reflects the perception that many users feel AI could develop beyond human control. Second, the group on economic dominance and power, such as "Oligarchic Capitalism and Wealth Inequality," "Economic Bubbles and Market Crashes," "Ponzi Schemes and Financial Fraud," "Sam Altman Public Perception Controversy," "Elon Musk and Tech Billionaire Scrutiny," and "Tech Elites and Technological Disruption," reflect concerns that the Stargate project will only strengthen the power of large corporations and tech elites while replacing human workers. Third, the environmental and infrastructure group, such as "Water Energy Nexus in Data Center," "Environmental Pollutions from Data Centers," "Data Center Development and Local Impact," and "Climate Change and Environmental Impact," reflects public concerns about the environmental impacts of data center development and their massive energy requirements.

On the other hand, positive sentiment was evident and appeared only in topics directly related to user experience or the quality of YouTube content as a source of information. The topic "Audience Receptions and Interview Content" is the only one with a dominant positive sentiment of 66%, indicating public appreciation for the quality of the interviews or the delivery of the information provided. Several other topics, such as "YouTube Investigative Journalism Channel" (49% positive), "Evaluation of Interview Quality and Performance" (42% positive), "Public Opinion on Presidential Leadership" with a 36% positive rate, "Evaluation of GPT Model Performance" with a 31% positive rate, and "ChatGPT Usage and User Interaction" with a 29% positive rate, also show slightly more positive views compared to other topics.

The findings of this study, which are dominated by negative sentiment, should not be interpreted solely as evidence of a negativity bias on the YouTube platform. A comparison with other studies [6] which combined two similar approaches BERTopic and BERT on the social media platform Reddit to analyze the future of AI, yielding predominantly positive sentiment. The difference in sentiment class dominance is primarily due to the difference in research subjects: one examines the future of AI in general, while the other focuses on a large-scale, prospective AI infrastructure project (Stargate) that is fraught with issues regarding corporate power concentration, environmental and economic impacts, and the figures behind it. The negative trends in this study are not uniform but depend on the characteristics of the topic. Positive sentiment dominates or is clearly evident in topics discussing the direct use of AI products, such as "ChatGPT Usage and User Interaction," "Evaluation of GPT Model Performance," as well as aspects of information dissemination or journalistic content such as "Investigative Journalism Content on YouTube," while topics with negative sentiment predominantly discuss the Stargate project prospectively, covering economic and environmental impacts, as well as the existential threat of AI in the future. This pattern aligns with research findings, where assessments of the direct use of AI products tend to be more positive. Thus, the dominance of negative sentiment found in this study may reflect the public's tendency to respond to discourse on large-scale projects or technologies that have not yet been fully realized with a more skeptical attitude compared to the use of technologies whose benefits have been directly experienced, as reflected in the difference in sentiment

distribution between topics regarding the prospective development of the Stargate project and topics regarding the direct use of AI in this dataset.

4. Conclusions and Future Works

This study successfully applied a method that integrates BERTopic topic modeling and RoBERTa sentiment analysis to examine public discussion topics related to OpenAI's Stargate project. The RoBERTa model was implemented by combining pseudo-labeling from a pre-trained model, an LLM-based annotation stage with Cohen's kappa calculation, class balancing, and the search for an optimal threshold to optimize accuracy, resulting in a model with an accuracy of 0.89 and a macro F1 score of 0.87. Meanwhile, the BERTopic model successfully identified 43 topics with a coherence score of 0.58 and a topic diversity of 0.71. The results of combining these two analyses revealed findings that address the research objectives, namely that the sentiment trends in public discussions regarding the OpenAI Stargate project are dominated by concerns, which include the existential threat of AI to human capabilities, economic power concentrated in the hands of the tech elite, and the environmental impact of expanding data center infrastructure. Meanwhile, the dominant topic with positive sentiment indicating public appreciation or hope arose from aspects of journalism and the information conveyed, as well as from direct experiences using AI technology, rather than from the Stargate project itself.

This study has some limitations that may open up opportunities for further research. First, this study analyzed data without considering the temporal dimension, so an approach to examine the evolution of public sentiment and discourse regarding the Stargate project over time has not yet been undertaken. Second, the data used is still limited to YouTube comments containing keywords representing OpenAI's Stargate project, so it does not yet cover public comments or discussions on other platforms that could potentially provide additional perspectives. Therefore, the limitations of this study can be addressed and expanded upon through future research.

5. Declaration

Author Contributions

Yuliana Dewi Proboningrum: Conceptualization, methodology, data collection, data analysis, and writing original draft. Hanifah Permatasari: Methodology review and supervision. Vihi Atina: Writing review.

Funding Statement

This research received no external funding.

Conflict of Interest

The authors declare no conflict of interest regarding this research and its publication.

Data Availability Statement

The dataset used in this study is publicly available in the authors GitHub repository. The source code developed for data collection, preprocessing, topic modeling, and sentiment analysis is not publicly available due to restrictions concerning the distribution of the research materials.

Use of Artificial Intelligence (AI)

The GPT based large language model (LLM) was used as an auxiliary tool to assist with the processing of the large volume of textual data in this study. In the sentiment analysis stage, the LLM was used to assist with the preliminary annotation of comments using a standardized prompt framework adopted from a previously published study. The LLM labels were subsequently reviewed by the author and their reliability was assessed through independent human annotation of a random sample using Cohen's kappa, yielding a value of 0.62, which indicated substantial agreement. Thus, the LLM outputs were treated as preliminary annotations and were subjected to human review and validation. In the BERTopic analysis, the LLM was used only after model evaluation to formulate descriptive topic labels based on the keywords generated by BERTopic, which were subsequently reviewed by the author for consistency with the identified topics. The LLM was not used to

determine the topics, evaluate the model, or make final scholarly judgements, which remained the responsibility of the author.

Ethical Clearance

This study used publicly available YouTube comments and did not involve direct interaction with human participants or the collection of personally identifiable information. Therefore, formal ethical approval was not required. The study was conducted in accordance with applicable ethical standards for the use and analysis of publicly available online data.

6. References

- [1] I. Zaenudin and A. B. Riyan, "Perkembangan Kecerdasan Buatan (AI) Dan Dampaknya Pada Dunia Teknologi," *Jitu: Jurnal Informatika Utama*, vol. 2, no. 2, pp. 128–153, 2024, doi: <https://doi.org/10.55903/jitu.v2i2.240>.
- [2] V. C. Storey, W. T. Yue, J. L. Zhao, and R. Lukyanenko, "Generative Artificial Intelligence: Evolving Technology, Growing Societal Impact, and Opportunities for Information Systems Research," *Information Systems Frontiers*, vol. 27, no. 5, pp. 2081–2102, Oct. 2025, doi: 10.1007/s10796-025-10581-7.
- [3] ClearML, AI Infrastructure Alliance, and Furiosa, "The State of AI Infrastructure at Scale 2024," 2024.
- [4] OpenAI, "Announcing The Stargate Project." Accessed: Feb. 20, 2026. [Online]. Available: <https://openai.com/id-ID/index/announcing-the-stargate-project/>
- [5] Bloomberg Originals, "Inside OpenAI's Stargate Megafactory with Sam Altman." Accessed: Jan. 30, 2026. [Online]. Available: <https://youtu.be/GhIJs4zbH0o?si=xN2be82cHkiAGMpn>
- [6] A. Ocal, "Perceptions of the Future of Artificial Intelligence on Social Media : A Topic Modeling and Sentiment Analysis Approach," *IEEE Access*, vol. 12, no. December, 2024, doi: 10.1109/ACCESS.2024.3510526.
- [7] W. Wahyuni, T. P. Lestari, M. Apriliana, and R. Gumelta, "Implementation of BERTopic for Topic Modeling Analysis of the Free Nutritious Meal Program Based on YouTube Comments," 2025.
- [8] B. Rahmadhani, Purwono, and S. D. Kurniawan, "Understanding Transformers: A Comprehensive Review," *Journal of Advanced Health Informatics Research (JAHIR)*, vol. 2, no. 2, pp. 85–94, 2024, doi: 10.59247/jahir.v2i2.292.
- [9] A. Söderwall and G. Telešova, "A Comparative Study with LDA and BERTopic : AI Policies Across Different Democracy Indexes," 2025.
- [10] M. David Angelo, R. Widyatna Harwenda, I. Budi, A. Budi Santoso, and P. Kresna Putra, "Sentiment Analysis and Topic Modeling of Public Opinion on Indonesia New Capital City Development Policies," *Journal Eduvest*, vol. 5, no. 5, pp. 5951–5969, 2025.
- [11] A. Alamsyah and N. D. Girawan, "Improving Clothing Product Quality and Reducing Waste Based on Consumer Review Using RoBERTa and BERTopic Language Model," *Big Data and Cognitive Computing*, vol. 7, 2023, doi: 10.3390/bdcc7040168.
- [12] A. Vishnevskaya, V. Pavlopoulos, and J. Stott, "Which Model Feels Better ? A Comparison of Computational Approaches to Emotion Detection in Social Media with Imbalanced Data," *Informatics*, pp. 1–24, 2026.

- [13] A. S. Aribowo, S. Khomsah, and S. Saifullah, "Semi-Supervised Sentiment Classification Using Self-Learning and Enhanced Co-Training," *JURNAL INFOTEL*, pp. 472–489, 2025, doi: <https://doi.org/10.20895/infotel.v17i3.1344>.
- [14] A. K. Jayaraman, A. Murugappan, T. E. Trueman, G. Ananthakrishnan, and A. Ghosh, "Imbalanced aspect categorization using bidirectional encoder representation from transformers," *Procedia Comput. Sci.*, vol. 218, pp. 757–765, 2022, doi: 10.1016/j.procs.2023.01.056.
- [15] L. Bojić *et al.*, "Comparing large Language models and human annotators in latent content analysis of sentiment , political leaning , emotional intensity and sarcasm." *Sci. Rep.*, pp. 1–16, 2025, doi: 10.1038/s41598-025-96508-3.
- [16] H. N. Alfiana, A. Doewes, and B. Widoyono, "Aspect-Based Sentiment Analysis of Access by KAI Application Reviews Using IndoBERT for Multi-Label Classification Tasks," *Jurnal Teknik Informatika (JUTIF)*, vol. 7, no. 1, pp. 286–306, 2026, doi: 10.52436/1.jutif.2026.7.2.5402 Aspect-Based.
- [17] W. J. Kusoema, I. Ibrahim, P. Studi, T. Informatika, K. Bandung, and J. Barat, "Analisis Sentimen dalam Kasus Korupsi PT . Pertamina menggunakan Metode indoBERT dan RCNN," *Sistemasi: Jurnal Sistem Informasi*, vol. 14, pp. 2246–2257, 2025.
- [18] S. Si, Y. Gao, H. Sun, Y. Zhang, and H. Luo, "Revisiting the Role of Label Smoothing in Enhanced Text Sentiment Classification," *Electronics*, pp. 1–15, 2026, doi: <https://doi.org/10.3390/electronics15101984>.
- [19] M. Abdelhamid and A. Desai, "Balancing the Scales : A Comprehensive Study on Tackling Class Imbalance in Binary Classification," 2024.
- [20] J. Hancock, J. M. Johnson, and T. M. Khosgoftaar, "A Comparative Approach to Threshold Optimization for Classifying Imbalanced Data," in *2022 IEEE 8th International Conference on Collaboration and Internet Computing (CIC)*, IEEE (Institute of Electrical and Electronics Engineers), 2022, pp. 135–142. doi: 10.1109/CIC56439.2022.00028.
- [21] N. Varshney, S. Mishra, and C. Baral, "Towards Improving Selective Prediction Ability of NLP Systems," in *7th Workshop on Representation Learning for NLP, RepL4NLP 2022 at ACL 2022*, Dublin: Association for Computational Linguistics (ACL), 2022, pp. 221–226.
- [22] B. D. Jesudas *et al.*, "Utilizing BERTopic Modeling for Concept Discovery in the Domain of Gerotranscendence and Solitude," *Res. Sq.*, pp. 1–23, 2025, doi: 10.21203/rs.3.rs-7383440/v1.
- [23] C. Schneijderberg, O. Wieczorek, and I. Steinhardt, *THE HANDBOOK OF QUALITATIVE AND QUANTITATIVE CONTENT ANALYSIS*. Taylor & Francis Group, 2026. doi: <https://doi.org/10.4324/9781003496397>.
- [24] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," Mar. 2022.
- [25] A. F. A. Mahmoud, F. A. Abdalla, G. Saad, M. Khamis, and M. A. Hamed, "A Comparative Evaluation of LDA, NMF, and BERTopic: Analyzing Perplexity and Coherence Metrics," *International Information and Engineering Technology Association*, vol. 30, no. 12, pp. 3163–3169, 2025.
- [26] A. Alharbi, A. A. Hai, R. Aljurbua, and Z. Obradovic, "AI-Driven Sentiment Trend Analysis: Enhancing Topic Modeling Interpretation with ChatGPT," *IFIP Adv. Inf. Commun. Technol.*, vol. 712, no. M1, pp. 3–17, 2024, doi: 10.1007/978-3-031-63215-0_1.
- [27] S. Samsir, R. S. Saragih, S. Subagio, R. Aditiya, and R. Watrianthos, "BERTopic Modeling of Natural Language Processing Abstracts: Thematic Structure and Trajectory," *Jurnal Media Informatika Budidarma*, vol. 7, no. 3, p. 1514, 2023, doi: 10.30865/mib.v7i3.6426.