Sentiments from Space: Insights from Space Experience Reviews with NLP

Nikshep Ash Kulli[1], Rohit Kumar Bandi RaviKumar[2], Rahul Prasad C[3],
Adjunct Professor, [2,3] Data Science Students
Department of Computer Science and Engineering,
Sacred Heart University, Fairfield, CT - 06825, USA.
{[1]kullin, [2]bandiravikumarr,[3]prasadr2}@mail.sacredheart.edu

Abstract— Space travel, an emerging frontier in tourism, offers unique experiences that captivate adventurous individuals. This paper presents an analysis of space travel experience reviews using advanced text processing and machine learning techniques. Text data was extracted from multiple websites, cleaned, and subjected to sentiment analysis and topic modeling. Sentiment analysis employed the VADER model to quantify review sentiments, while Latent Dirichlet Allocation (LDA) was used for topic modeling to identify prevalent themes. Visualizations such as word clouds and scatter plots were created to illustrate the findings. The results reveal key insights into customer satisfaction and common concerns, providing valuable information for stakeholders in the space tourism industry.

Index Terms - NLP, sentiment analysis, topic modeling, space tourism, LDA, VADER.

I. Introduction

In the modern era of space exploration, space travel has become a captivating topic for adventurers and technology enthusiasts alike. As the commercialization of space tourism takes off, understanding customer feedback becomes crucial for improving services and ensuring satisfaction. Analyzing reviews of space travel experiences provides valuable insights into customer sentiments and expectations.

Given the complexity and novelty of space travel, customers often share detailed experiences and feedback online. This research aims to consolidate and analyze these reviews using advanced text processing and machine learning techniques. By employing sentiment analysis and topic modeling, we can uncover prevalent themes and sentiments expressed by customers, offering a comprehensive understanding of their experiences.

Sentiment analysis helps quantify the emotional tone of the reviews, while topic modeling identifies the main themes discussed. These insights are vital for stakeholders in the space tourism industry to make informed decisions and enhance service quality. This study leverages natural language processing (NLP) tools and visualization techniques to present the findings in an intuitive manner, making the feedback easily interpretable and actionable.

The article has different sections. Section II provides details about state of the art literature review carried out. Section III briefs about the proposed approach for text analysis. Section IV provides insight to the obtained visualization and the analysis part. Section V concludes the work with scope for future work.

II. Literature Survey

Akash Sharma et al. (2024), in their study focuses on leveraging machine learning techniques such as text preprocessing, feature engineering, and sentiment analysis to categorize customer reviews into three categories: neutral, positive, and negative. By employing algorithms like Support Vector Machine (SVM) and Naïve Bayes, the research aims to automate the analysis of sentiment from Amazon's e-commerce reviews. This approach not only enhances the efficiency of understanding customer sentiments at scale but also facilitates informed decision-making for product improvements and marketing strategies [1].

Fiqui Amali et al. (2024), this research focuses on analyzing customer feedback. Advanced deep learning techniques are employed to extract nuanced sentiments from a large dataset of hotel reviews. Deep neural networks such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks are utilized for sentiment analysis. The study compares various network architectures including CNN, LSTM, CNN-LSTM, BiLSTM, and ConvBiLSTM to identify the most effective approach. Experimental results using datasets from Indonesian hotels crawled from TripAdvisor show LSTM achieving 96.42% accuracy on the Padma Hotel dataset and 85.31% on the Hard Rock Hotel dataset. The CNN-LSTM model achieves 85.87% accuracy on the Ayana Hotel dataset, while BiLSTM achieves 86.09% on the Pullman Hotel dataset. Comparative analysis against commonly used IMDb datasets highlights the superior performance of deep learning models over traditional machine learning models across all metrics [2].

Lulu Fan et al. (2024), this study proposes a novel approach combining sentiment lexicons and machine learning to construct a tourism emotion model. Initially, tourist texts from micro-blogs and travel websites are used to build a tourism emotional lexicon. Sentences with clear emotional polarity are extracted based on sentiment scores, forming an initial classifier using the Naïve Bayes (NB) algorithm. The sentiment classification model is further refined using the Latent Dirichlet Allocation (LDA) topic model. This refined model significantly improves classification accuracy compared to both the sentiment lexicon method and machine learning alone. Experimental results on Lijiang travel texts demonstrate a 13.85% increase in accuracy over the sentiment lexicon method and an 8.51% increase over standalone machine learning methods [3].

Haiyan Zhan et al. (2022), In order to effectively mine hidden research topics and potential evolution patterns from massive network public opinion data, this paper proposes a public opinion monitoring system based on LDA. The scheme mainly performs text crawling and topic extraction of network entities through self-programming by Python and realizes fine-grained emotion analysis and topic mining combined with TF-IDF feature words and LDA model. Then, cluster analysis is made on the content and intensity of each stage of public opinion development, and the construction of topic model and topic confusion are extracted for further public opinion prediction. Finally, the LDA topic model is applied to the application of public opinion system, which realizes

the intelligent monitoring, analysis and prediction function of public opinion, and it can quickly respond to the update and change of network public opinion [4].

Dan Wang et al. (2021), This paper utilizes data crawling to collect information and reviews of popular movies. Through data visualization and LDA theme analysis, the study identifies general trends in movie development and audience preferences. The findings indicate that China currently dominates much of the global movie market. However, movie industry professionals must focus on creating realistic, high-quality films that align with audience needs to sustain and enhance this growth [5].

Chu Zhang et al. (2022), proposes a method to extract users' STS requirements using an improved TF-IDF and sentiment analysis approach. Initially, user keywords are identified through text segmentation, followed by the calculation of user requirement weights using TF-IDF. To enhance accuracy, Word2Vec is employed to filter synonyms, improving the TF-IDF process. Finally, sentiment analysis is applied to adjust the weights of user requirements. Experimental results demonstrate the efficiency of this method [6].

Ram Krishn Mishra et al(2019), The rapid evolution of the internet has led many users to rely on online services for daily tasks. The tourism industry significantly benefits internet users by offering economical and comparable prices for hotel bookings. Nowadays, many users share their feedback through blogs, web forums, social media, and other platforms. This research employs term frequency-inverse document frequency (TF-IDF) and cosine similarity to suggest additional hotel options based on user reviews. TF-IDF is used to determine the weight of terms or document frequency, while cosine similarity extracts similar values from the sentiment dataset [7].

III. Proposed Methodology

The proposed system focuses on analyzing space travel experience reviews using advanced text processing and machine learning techniques. The main features of the system are as follows:

- Extracting and preprocessing text data from different websites.
- Performing sentiment analysis on the reviews.
- Conducting topic modeling to identify prevalent themes.

Step 1: Data Extraction

The system extracts text data from multiple websites containing space travel reviews using web scraping. This process begins with identifying relevant websites, such as travel forums, review platforms, and blogs, that host substantial user-generated content on space travel experiences. Web scraping tools like BeautifulSoup, Scrapy, and Selenium are employed to automate the extraction of HTML content from these sites. Using HTTP requests, the system retrieves web pages and parses the HTML to locate review sections. BeautifulSoup helps navigate and extract specific elements from the HTML structure, while Selenium handles dynamic content by simulating user interactions to load hidden elements. This ensures comprehensive data collection, capturing reviews that might be behind interactive features.

Once the HTML content is extracted, it undergoes thorough cleaning to remove irrelevant elements like advertisements and navigation menus. This involves using regular expressions and string manipulation techniques to isolate meaningful content. The cleaned data is then converted into a structured format, typically a CSV file or a database, with organized columns for review text, author names, dates, and ratings. This structured data format facilitates subsequent analysis steps, such as sentiment analysis and topic modeling, by ensuring the information is uniformly organized and easily accessible. This meticulous approach to data extraction and structuring is crucial for obtaining high-quality, analyzable datasets from diverse web sources.

Step 2: Data Pre-Processing

Data preprocessing is essential before conducting analysis or using machine learning models. This process involves several steps to clean and prepare the extracted data. Initially, numbers and special characters are removed using regular expressions (regex). For instance, patterns like \d+ match digits, while \W+ matches non-word characters, ensuring the text is devoid of numeric and non-alphanumeric symbols.

Next, short words are eliminated to focus on more meaningful content. Natural Language Processing (NLP) techniques are applied here. After tokenizing the text into individual words, tokens shorter than a specified length threshold (e.g., less than three characters) are filtered out. This step helps in reducing noise and irrelevant details that may not contribute significantly to the text's meaning.

Additionally, stopwords are removed during preprocessing to further refine the text. Stopwords are common words such as "the", "is", and "and", which are typically filtered out because they do not carry significant semantic meaning. This step involves comparing tokenized words against predefined lists of stopwords for the language of the text, ensuring that the analysis focuses on words that convey important information.

Step 3: Sentimental Analysis

The provided data undergoes sentiment analysis using the VADER (Valence Aware Dictionary and sEntiment Reasoner) model, which is designed for sentiment analysis of social media text. The analyze_sentiment function first initializes the SentimentIntensityAnalyzer object from the VADER library. It then calculates the sentiment scores for the input text, returning a dictionary with various sentiment metrics: positive, negative, neutral, and a composite (compound) score. The compound score is a normalized, weighted aggregate of the other scores, ranging from -1 (most negative) to 1 (most positive). Based on the compound score, the sentiment is classified into three categories: 'Positive' (compound score ≥ 0.05), 'Negative' (compound score ≤ -0.05), and 'Neutral' (compound score between -0.05 and 0.05). This function is applied to each review in the 'cleaned_reviews' column of the DataFrame, and the resulting sentiment classification is stored in a new column called 'sentiment'.

Once the sentiment analysis is complete, the code calculates the distribution of sentiment categories within the reviews. By using the value_counts method with normalization, it determines the percentage of reviews that fall into each sentiment category ('Positive', 'Negative', and 'Neutral'). For instance, the output shows that 27.68% of the reviews are positive, while 11.81% are negative. These percentages provide an overall view of customer sentiment. To visualize this distribution, a pie chart is created using Matplotlib. The chart is designed with distinct colors for each sentiment category (green for positive, red for negative, and grey for neutral) and includes percentage labels. The visualization

offers an intuitive and immediate understanding of the sentiment breakdown, highlighting the relative proportions of positive, negative, and neutral reviews.

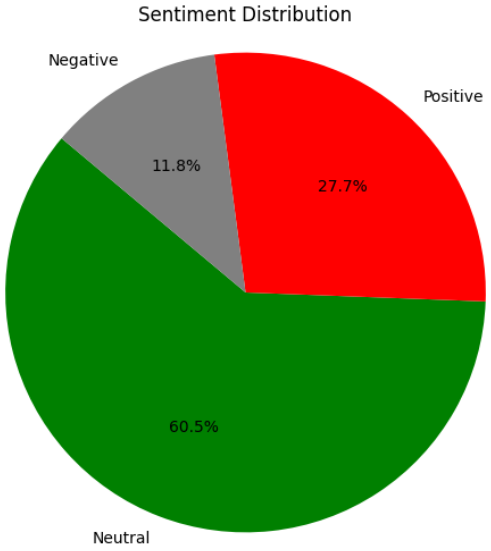

Figure 1 – Sentimental Analysis of Reviews

Step 4: TF-IDF Analysis Process

The process of Term Frequency-Inverse Document Frequency (TF-IDF) analysis begins with transforming the cleaned reviews into a numerical format that captures the importance of each word in the context of the entire dataset. To achieve this, the TF-IDF vectorizer is initialized with a limit of 1,000 features and configured to ignore common English stop words. The vectorizer is then applied to the 'cleaned_reviews' column, creating a TF-IDF matrix. This matrix represents the importance of words in each document, where each row corresponds to a document (review), and each column corresponds to a term (word). The values in the matrix are the TF-IDF scores, which quantify the significance of each word in the document relative to the entire corpus.

After generating the TF-IDF matrix, the next step involves calculating the TF-IDF scores for each document. For each review, the non-zero entries in its corresponding row of the TF-IDF matrix are identified. These entries indicate the terms present in the review and their associated TF-IDF scores. By mapping these scores to their respective terms using the feature names obtained from the vectorizer, a dictionary of terms and their scores is created for each document. These dictionaries are then sorted in descending order of the TF-IDF scores to highlight the most important terms for each review. This sorting helps identify the top terms that are most representative of each document based on their TF-IDF scores.

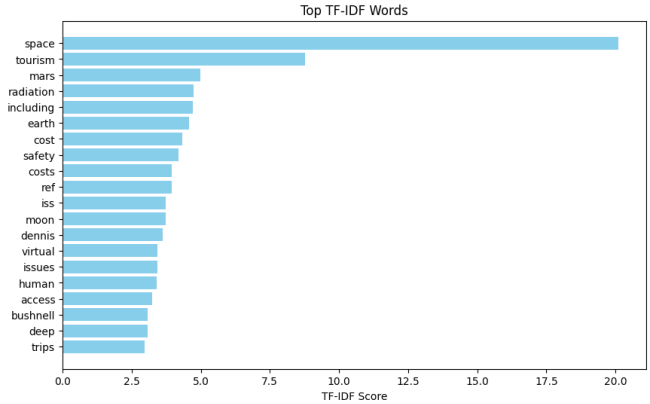

Figure 2– Top TF-IDF Words

To provide a clearer understanding of the TF-IDF scores, the top terms for each document are printed along with their scores. This highlights the most significant terms for each review, offering insights into the key themes and topics discussed by reviewers. Additionally, the TF-IDF scores for a specific document can be visualized using a bar chart. For example, selecting a document and plotting its top 10 terms with the highest TF-IDF scores creates a visual representation that makes it easy to see which terms are most important in that document. The bar chart shows the terms on the y-axis and their corresponding TF-IDF scores on the x-axis, with the bars sorted in descending order. This visualization not only underscores the importance of specific terms but also provides an intuitive way to interpret the results of the TF-IDF analysis

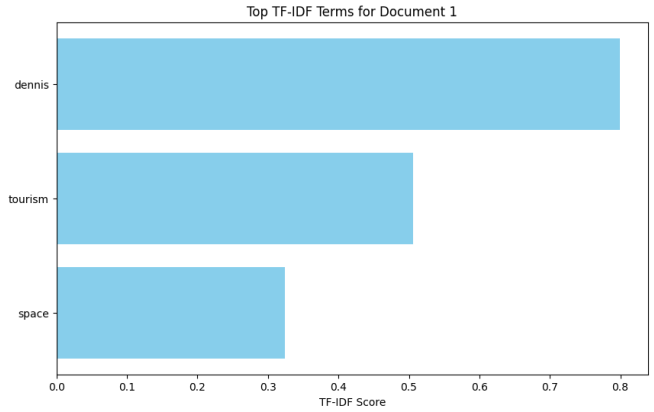

Figure 3– TF-IDF for Document 1

Step 5: Topic Modeling (LDA)

Topic modeling is employed to uncover hidden themes within a set of documents. In this case, Latent Dirichlet Allocation (LDA) is utilized to identify topics in the cleaned reviews. Initially, an LDA model is initialized with a predefined number of topics (in this example, eight). The model is then fitted to the TF-IDF matrix, which contains the numerical representations of the reviews. Fitting the LDA model to the TF-IDF matrix involves discovering the underlying structure in the data and grouping the terms into distinct topics based on their co-occurrence patterns within the documents. Each topic is represented as a distribution over the terms, and each document is represented as a distribution over the topics.

To interpret the topics identified by the LDA model, the most significant words for each topic are displayed. This is done using a function that extracts the top words for each topic from the model's components. For each topic, the words with the highest weights are printed, providing a clear picture of the key terms that define that topic. For instance, the function prints out the top five words for each of the eight topics, allowing us to infer the main themes discussed in the reviews.

To further aid in understanding, the topics are visualized using bar charts. Each topic's top words are plotted, with their weights indicating the importance of each word within the topic. The bar charts are arranged vertically, one for each topic, making it easy to compare and contrast the topics. The bars are color-coded and sorted in descending order, with the most significant words at the top. This visualization helps to intuitively grasp the essence of each topic, highlighting the dominant terms that contribute to the thematic structure identified by the LDA model. The combination of displaying and visualizing the topics offers a comprehensive understanding of the latent themes present in the reviews.

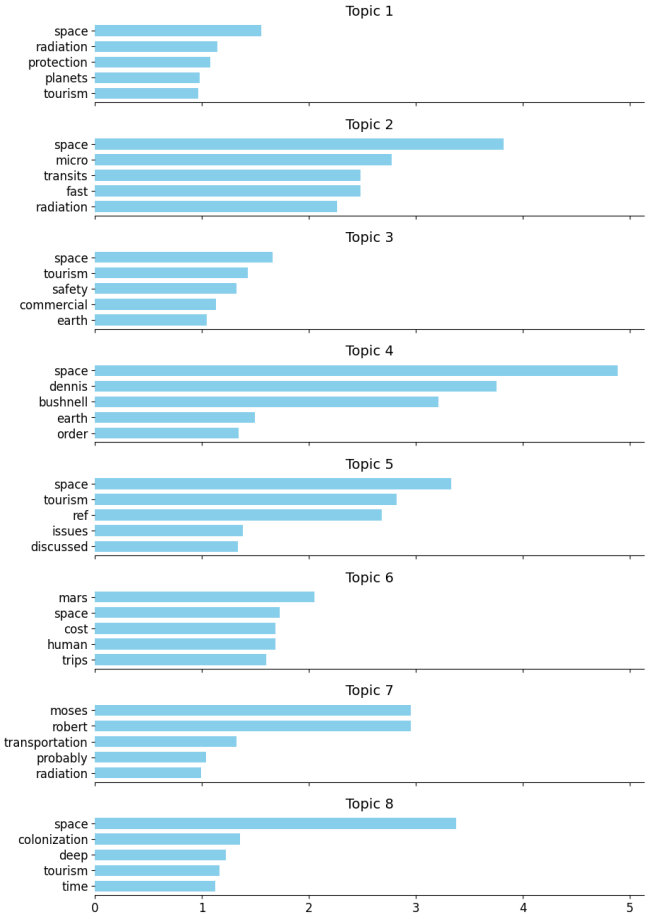

Figure 4– Topic Modeling

IV. Visualization

In this section, various visualizations are created to better understand the textual data and its sentiment. These visualizations include different forms of word clouds and a scatter plot for sentiment analysis.

1. General Word Cloud
   - Purpose: To quickly identify the most frequent words and themes in the dataset.
   - Benefits: Provides a high-level overview of the data, helping to pinpoint common topics and key terms.

A general word cloud is a visual representation of the most frequently occurring words in a corpus of text, in this case, the cleaned reviews. In this visualization, the size of each word indicates its frequency or importance: larger words appear more frequently in the dataset. This type of visualization is useful for quickly identifying the main topics or themes discussed in the reviews. By generating a word cloud for all documents, we can see at a glance which words dominate the discussion, providing insights into what aspects of the reviews are most prominent or recurring.

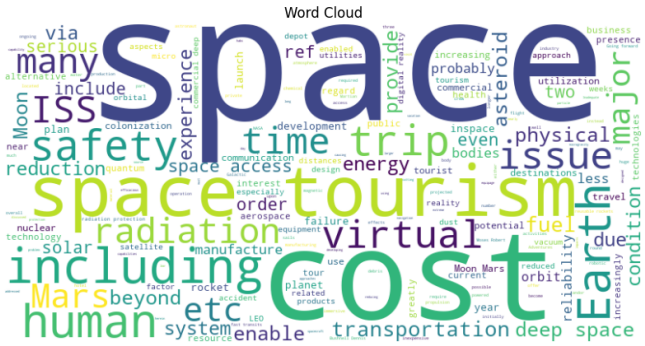

Figure 5– Word Cloud

2. Space Rocket-shaped Word Cloud
   - Purpose: To present key terms in a visually engaging and contextually relevant manner.
   - Benefits: Enhances thematic relevance, making the visualization more engaging and suitable for presentations related to space travel.

Creating a word cloud in the shape of a space rocket adds a thematic and visually engaging element to the visualization. This is done by using a mask image of a space rocket. The word cloud generator arranges the words within the confines of the rocket shape, maintaining the relevance of word sizes to their frequency. This method not only highlights the key terms but also ties the visualization to the context of space travel, making it more appealing and contextually appropriate for presentations or reports related to space travel experiences.

Figure 6– Space-Rocket Shaped Word Cloud

3. Astronaut-shaped Word Cloud
   - Purpose: Similar to the rocket-shaped word cloud, it aims to present key terms within a thematic and visually appealing format.
   - Benefits: Reinforces the theme of space exploration, making the insights more engaging and memorable.

Similar to the rocket-shaped word cloud, an astronaut-shaped word cloud uses a mask image of an astronaut. This visualization method enhances the thematic connection to space exploration. Words are arranged within the shape of an astronaut, with larger words indicating higher frequency. Such visualizations can be particularly engaging for audiences interested in space travel, as they combine informative content with a visually appealing format. The astronaut shape reinforces the theme and can make the insights drawn from the text data more memorable.

Figure 7– Astronaut-shaped Word Cloud

4. Sentiment Analysis Visualization

- Purpose: To visualize the distribution and trends of sentiment across the dataset.
- Benefits: Helps in understanding the overall sentiment, identifying trends, patterns, and outliers, and providing a comprehensive view of the emotional tone in the reviews.

Sentiment analysis aims to quantify the sentiment expressed in text, categorizing it as positive, negative, or neutral. In this visualization, a scatter plot is used to represent the sentiment scores of each review. The x-axis typically represents the document index (or review number), while the y-axis shows the sentiment score. Each point on the scatter plot corresponds to a single review, with its position on the y-axis indicating its sentiment score (positive, negative, or neutral). This type of visualization helps in identifying the overall sentiment trend in the dataset. It can reveal patterns such as clusters of positive or negative sentiment, indicating common feelings or opinions among the reviews. Additionally, outliers can be easily spotted, which might warrant further investigation.

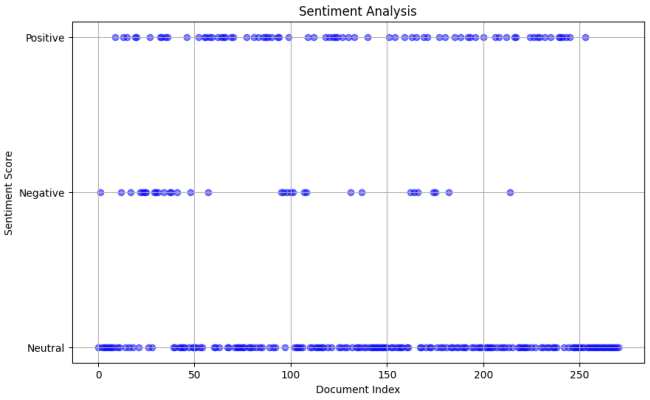

Figure 8– Sentimental Analysis of Reviews

V. Conclusion

This paper presents a thorough analysis of customer reviews related to space travel experiences, employing a range of natural language processing (NLP) techniques to extract meaningful insights. The process began with data cleaning to prepare the reviews for analysis, followed by sentiment analysis using the VADER model. This allowed us to categorize the reviews as positive, negative, or neutral, revealing that 27.68% of the reviews were positive and 11.81% were negative. We then applied TF-IDF to identify significant terms and conducted topic modeling with LDA to uncover eight distinct topics, each represented by a set of top words. These analyses provided a deeper understanding of the key themes and sentiments expressed by customers.

To enhance the interpretability and engagement of our findings, we employed various visualization techniques. A general word cloud highlighted the most frequent terms, while thematic word clouds in the shapes of a space rocket and an astronaut added a contextual and visually appealing dimension. Additionally, a scatter plot visualized the distribution of sentiment scores across the dataset. These visualizations collectively offered a comprehensive and engaging way to explore and understand the text data, highlighting key terms and sentiment trends while maintaining a strong thematic connection to space travel. This multifaceted approach provides valuable insights into customer experiences and perceptions in the burgeoning field of space travel.

VI. Refferences

[1] Gitanshu Chauhan, Akash Sharma, Nripendra Dwivedi, “Amazon Product Reviews Sentimental Analysis using Machine Learning”, in 2024 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT), 09-10 February 2024.

[2] Fiqui Amali, Halil Yigit, Zeynep Hilal Kilimci, “Sentiment Analysis of Hotel Reviews using Deep Learning Approaches”, in 2024 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream), 25-25 April 2024.

[3] Lulu Fan, Bingyan Chen, Xiabao Fu, “Sentiment Classification of Tourism Based on Rules and LDA Topic Model”, in 2019 International Conference on Electronic Engineering and Informatics (EEI)”, 08-10 November 2019.

[4] Haiyan Zhan, “Network Public Opinion Analysis Based LDA Model”, in 2022 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS), 26-27 March 2022.

[5] Zhiqu Li, Dan Wang, Yitong Wu, “Sentiment Analysis on Chinese Movie Comment with LDA Model”, in 2021 2nd International Conference on Big Data Economy and Information Management (BDEIM), 03-05 December 2021.

[6] Yushan Xu, Chu Zhang, Wenyan Song, “Prioritizing Customer Requirements for Science and Technology Service Platform Based on Improved TF-IDF and Sentiment Analysis”, in 2022 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), 07-10 December 2022.

[7] Ram Krishn Mishra, Siddhaling Urolagin, “A Sentiment analysis-based hotel recommendation using TF-IDF Approach” in 2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE), 11-12 December 2019.

[8] Yeresime Suresh, Br Rohit Kumar, C Jahnavi Reddy, P Sai Rohini, K Sharath , “Naive Bayes Classifier based Movie Recommendation System”, in 2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS),14-16 June 2023.

[9] Muhammad Alkaff, Andreyan Rizky Baskara, Yohanes Hendro Wicaksono, "Sentiment Analysis of Indonesian Movie Trailer on YouTube Using Delta TF-IDF and SVM", in 2020 Fifth International Conference on Informatics and Computing (ICIC), 03-04 November 2020.

[10] Ashwani Gupta, Utpal Sharma, "Machine Learning based Sentiment Analysis of Hindi Data with TF-IDF and Count Vectorization", in 2022 7th International Conference on Computing, Communication and Security (ICCCS), 03-05 November 2022.