DECODING REBAL D'S YOUTUBE STRATEGY: HOW TOPICS AND SENTIMENT SHAPE CONTENT APPEAL

ABSTRACT. Many studies provide pieces of advice for expanding a YouTube channel, yet each YouTuber distinguishes themselves through distinct content tactics. This research investigates the qualities of a successful YouTuber by examining Rebal D's case, a content creator on YouTube with a young target audience. By collecting data from his YouTube channel, we evaluate significant metrics such as views, likes, and comments, as well as employ topic modeling, sentiment analysis, and network analysis, to identify trends in his content style and tactics for engaging his audience. This study intends to inform content creators and marketers on effective engagement tactics that are suited to the enhancement of audience engagement on social media.

Keywords: YouTube; Sentiment Analysis; Topic Modeling; Content Strategy; Emotions

INTRODUCTION

YouTube has become one of the most influential platforms in the digital age, amassing over 2 billion monthly users globally. The platform's reach allows content creators to engage with diverse audiences, leveraging various tools to captivate viewers and maintain sustained interest [1]. Research has consistently shown that content creation, topic choice, and emotional appeals significantly impact user engagement on platforms like YouTube. However, while numerous studies have explored general social media engagement strategies, a gap exists in understanding how specific creators craft their content to achieve audience retention and interaction [2].

This study focuses on Rebal D, a renowned content creator on YouTube recognized for his witty and sarcastic perspective on social media trends. Rebal D, who has over 2.8 million followers, engages his viewers with humor, social criticism, sarcasm, and interactions. However, the particular content tactics that contributed to his success remain unexplored. This study tries to fill that void by looking into the thematic and emotional components of his videos, as well as how they contribute to viewer engagement.

This study will uncover reoccurring topics and emotional tones that are important to Rebal D's content strategy by examining video titles and descriptions using topic modeling and sentiment analysis. Furthermore, using network analysis, we will investigate how these topics are interrelated and contribute to the overall structure of his video material. The study uses approaches including Latent Dirichlet Allocation (LDA) [3] for topic modeling and the VADER sentiment analysis tool [4] to identify emotional patterns.

The goal of this study is to identify trends in Rebal D's content that may help him interact with his audience more successfully. This research provides a deeper knowledge of the link between topics, emotional tone, and audience engagement. It may influence tactics for both content providers and marketers to optimize viewer retention and loyalty on digital platforms [5].

In the next section, we will start by discussing the methodologies employed, including the specific techniques used to analyze video content and uncover these patterns.

METHODOLOGY

1.1 Data Collection

Since this study is unprecedented, the data retrieval had to be done promptly using web scraping techniques from Rebal D's YouTube channel. This allowed for the programmatic access of structured data from web pages, which can be especially useful when analyzing online content.

- **Information Extracted**
  Fetching and parsing of the HTML content allowed us to extract the following information: **Channel title, Video description, Publish date, View count, Like count, Comment count.**

- **Retrieval of Video Data**
  The retrieval of video data from Rebal D's YouTube channel was accomplished through the use of the **YouTube Data API**. This approach facilitates the extraction of detailed information about each video, enabling comprehensive analysis of the channel's content.

1.2 Data Manipulation

This part is dedicated to data cleaning and preprocessing, as well as data exploration and visualization. The cleaning methods employed enhance the integrity of the data, thereby allowing for more accurate insights into viewer engagement and content strategy.

1.2.1 Data cleaning and Preprocessing

In this stage of the research, we focused on cleaning and preprocessing the video data to ensure accuracy and relevance in our analysis.

- **Converting Columns**
  Initially, numerical columns like "view count," "like count," and "comment count" were converted to integer format to enable accurate analysis and comparisons. The "publication date" was also converted to date time format for effective temporal analysis.

- **Pre-processing Video Descriptions**
  To ensure the descriptions contain only relevant information, we remove unwanted content, such as promotional material: sponsorship, irrelevant calls to action, and advertisements, e.g. "Big Thanks to '...' for sponsoring this video", as these could bias the analysis towards external marketing strategies rather than content engagement.

- **List of Unwanted Phrases**
  We manually compiled a comprehensive list of specific phrases and links that were deemed unnecessary for analysis. This list served as an additional filter to ensure a cleaner output.

1.2.2 Data exploration and visualization

Data exploration was conducted to get an in-depth understanding of the data we are working with. First, we visualized the **distribution of views, likes, and comments** in the videos, as well as the **trend of video publications over time** to assess the frequency and consistency of video publications. Then we visualized the **lengths of titles and descriptions** to explore any recurring patterns or strategies in Rebal D's choice of title and description lengths, assessing their role in content presentation. Last but not least, we ran a quick **correlation study** of the relationships between the variables indicated in the data showcased by a correlation matrix.

1.3 Topic Modeling Using LDA

Topic modeling is a technique used to identify hidden themes within textual data by grouping together words that frequently appear in the same context to reveal the underlying topics present in data. The goal of this analysis is to uncover the dominant themes in the YouTube video descriptions and better understand the type of content being produced.

In this study, we applied **Latent Dirichlet Allocation (LDA)** [3], which is one of the most widely used methods for topic modeling, to group similar words from the YouTube descriptions to identify distinct topics within the content. This provides valuable insights into Rebal D's thematic trends, allowing us to interpret how these themes align with audience engagement strategies.

1.3.1 Preprocessing Specific to this Step

Before applying LDA, the text data underwent several preprocessing steps to improve the quality of the analysis:

- **Lowercasing:** To ensure consistency across all text entries.
- **Number Removal:** Numbers often do not contribute meaningful information to the thematic structure.
- **Punctuation Removal:** Punctuation marks were removed to avoid interference in word tokenization.
- **Tokenization:** Each description was broken down into individual words or tokens.
- **Removing Short Words:** Words with fewer than three characters were excluded to filter out noise.
- **Stopwords Removal:** Common, non-informative words like "the" and "and" were removed using standard stopword lists.
- **Lemmatization** Words were reduced to their root forms, ensuring variations of the same word are treated as one (e.g., "girls" becomes "girl").

1.3.2 Choosing the Optimal Number of Topics

To determine the best number of topics, multiple LDA models with varying topic counts were tested. The selection was based on a **coherence score**, which measures how **interpretable and meaningful** the topics are. After examining different models, four topics were selected, as this provided a good balance between **coherence** and **thematic richness**.

1.4 Network Analysis

In the network analysis conducted for this study, the primary aim was to explore how videos relate to the various topics identified through LDA modeling. To construct the network, a **NetworkX** [6] graph was employed, with nodes corresponding to videos (represented by their respective indices) and topics derived from LDA modeling. The edges between these nodes were used to represent thematic connections. By plotting this graph, the resulting visualization offers a clear, structured view of how videos align with different themes. Video nodes were labeled by index numbers to enhance clarity, while edges were made semi-transparent to reduce visual clutter, improving the graph's readability.

1.4.1 Centrality Metrics

Several **centrality metrics** were calculated to further interpret the importance of individual videos and topics within the network:

- **Betweenness Centrality** quantifies the number of times a node acts as a bridge along the shortest path between two other nodes. [7]
- **Degree Centrality** counts the number of edges (or direct connections) that a node has, providing insights into its direct influence within the network. [8]
- **Eigenvector Centrality** assesses the influence of a node in terms of its connections to other influential nodes. [9]

These metrics provide a comprehensive view of the thematic structure in Rebal D's videos, revealing which content serves as **thematic keystones** within his channel and how these keystones may drive viewer engagement across multiple content areas.

1.5 Sentiment Analysis

In our sentiment analysis, we split the methodology into two approaches: one using the VADER sentiment analysis tool, and the other leveraging a pre-trained RoBERTa-based model via HuggingFace for emotion detection.

1.5.1 VADER sentiment analysis

The first part of the sentiment analysis employs the **VADER (Valence Aware Dictionary and sEntiment Reasoner)** tool [10]. This tool classifies text into three primary sentiments: positive, neutral, and negative, and assigns a sentiment score that indicates the intensity of these sentiments. VADER is widely recognized for its effectiveness in handling the informal, emotive, and often non-standard language used in social media contexts, making it especially suitable for YouTube video content analysis.

1.5.2 Emotion Detection with RoBERTa

For a deeper analysis, we extend beyond traditional sentiment classification and explore emotional nuance using a pre-trained RoBERTa-based model for emotion detection [11]. This model can classify text into multiple emotional categories, such as anger, joy, sadness, fear, surprise, and love, which offer more granularity than the VADER sentiment scores.

Both methods allow us to evaluate not only the overt sentiment but also the underlying emotional currents in Rebal D's video titles and descriptions.

RESULTS AND FINDINGS

2.1 Data Cleaning and Preprocessing

The summary statistics for the data's numerical columns provided insights into the performance metrics of Rebal D's videos. On average, each video gains on average **1,033,327 views, 67,123 likes, and 3,623 comments**. The highest recorded metrics were notable, with a maximum of **6,399,852 views, 281,229 likes, and 20,975 comments**.

2.2 Data Visualization

2.2.1 Distribution of views, likes, and comments

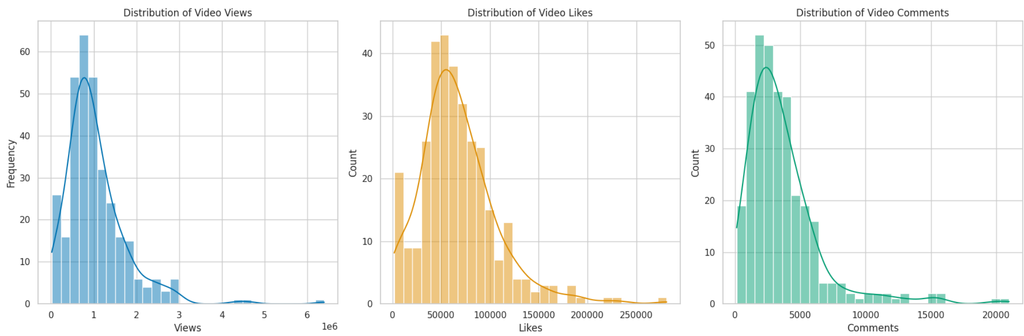

Figure 1. Distribution of Views, Likes and Comments

The distributions of the three metrics—views, likes, and comments— (Figure 1) exhibit **right-skewed** patterns, meaning that the mean is significantly higher than the median. This skewness is attributed to the presence of outliers, which are a small number of exceptionally high-performing videos that inflate the mean without accurately representing typical video performance.

2.2.2 Trend of Video Publications Over Time

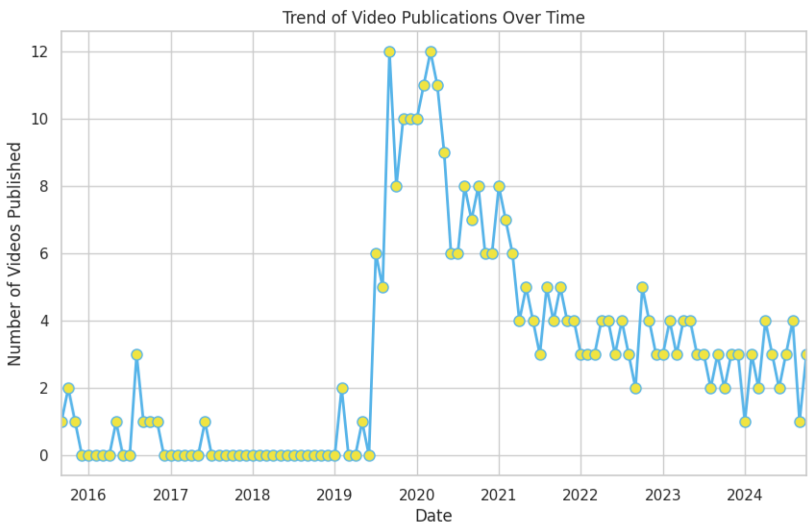

Figure 2. Trend of Video Publication Over Time

According to Figure 2, in the early years of his career, Rebal D demonstrated inconsistent video production, with only **13 videos** released between late 2015 and early 2019. It was not until 2019 that his productivity significantly increased, peaking at approximately **64 videos** per year from 2019 to 2020. Following this peak, there was a gradual decline in output, stabilizing at around **38 videos** per year starting in 2021.

2.2.3 Titles and Description lengths

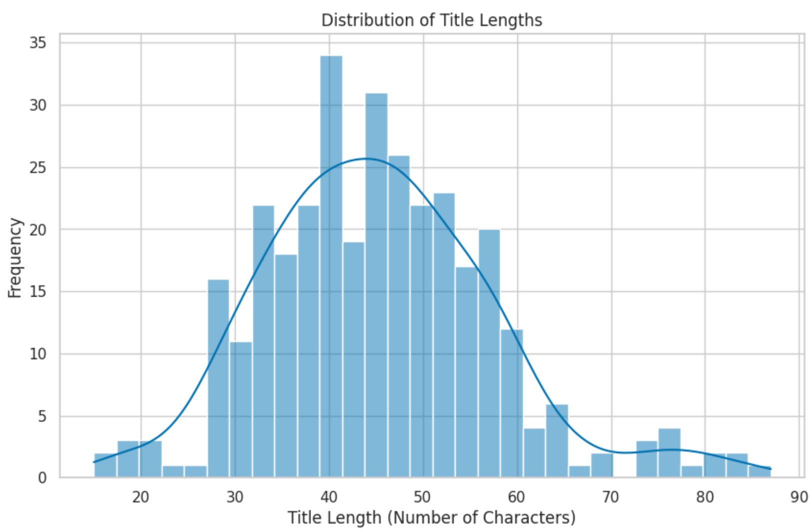

Figure 3. Distribution of Description lengths

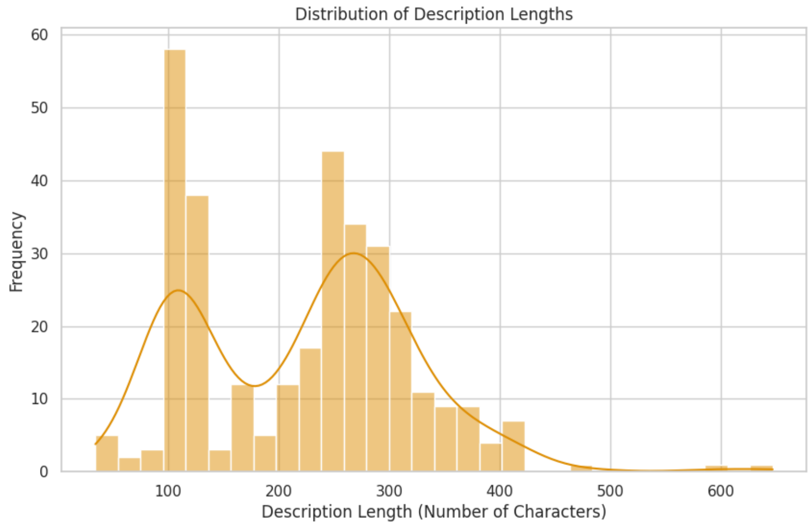

Figure 4. Distribution of Description lengths

Figure 3 appears normal (bell-shaped), suggesting a **typical range of title lengths** that resonates well with the audience. This pattern indicates a **consistency** in how titles are crafted across videos. In contrast, Figure 4 shows a bimodal pattern, indicating **two prevalent lengths** for video descriptions (120 characters, and 260 characters). This suggests varying strategies in crafting descriptions, with some being concise and others more elaborate.

2.2.4 Correlation between variables

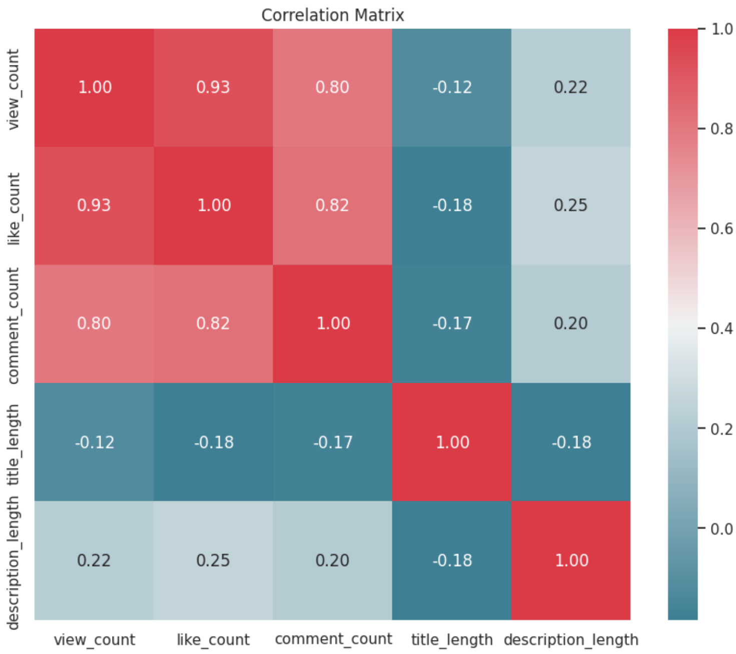

Figure 5. Correlation Matrix Between Variables

The correlation matrix (Figure 5) between the variables of the study indicates a **strong positive relationship** between "comment count", "like count", and "view count", suggesting that as one metric increases, the others also tend to increase. However, the weak correlations (around 0.2 to -0.18) between these metrics and title and description lengths indicate that **length alone does not significantly impact audience engagement.** This highlights the need for further analyses to uncover deeper relationships within the data.

2.3 Topic Modeling Using LDA

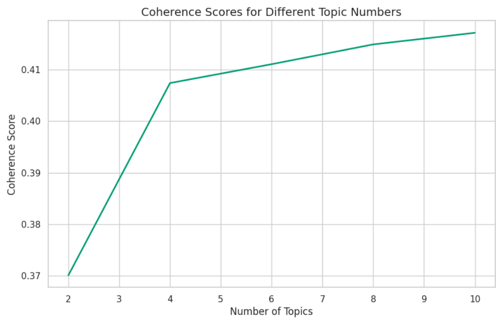

Figure 6. Coherence Scores for different thematics

The following Coherence Scores graph (Figure 6) shows that having more topics resulted in slightly higher coherence. However, increasing the number of topics beyond 4 caused significant **overlap**, especially between themes related to TikTok and personal commentary, hence, reducing the number of **topics to 4**.

- **Topic 0** focuses on TikTok content and trends, especially around gender dynamics and social media challenges.
- **Topic 1** is centered around engagement tools and calls to action (subscribe, merch, join).
- **Topic 2** captures content around dating, awkward social situations, and reactions, blending reality-TV-like scenarios.
- **Topic 3** reflects broader reflections on people, life, and commentary on personal relationships.

2.4 Network Analysis

The network graph (Figure 7) presents nodes representing individual videos and the topics associated with them.

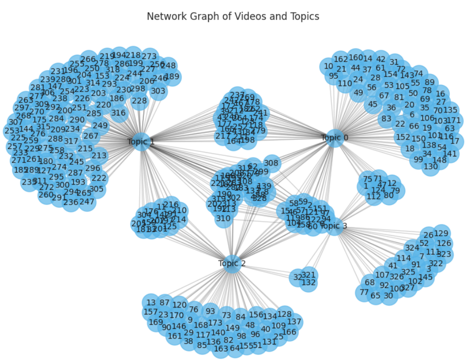

Figure 7. Network Graph of Videos and Topics

Each edge between a video and a topic symbolizes a thematic link, effectively mapping how a video fits within one or more of the topics identified through the LDA model.

Topic Associations: Many videos were linked to multiple topics, indicating a rich thematic overlap within the content. This multi-topic association suggests that videos often engage with several themes, reflecting the **complexity of audience interests.** **Topic 1**, which has the highest number of nodes and edges, stands out due to its numerous shared connections with other topics. This topic, **centered on calls to action**, highlights the critical role of engagement tools in shaping a YouTuber's strategy.

In addition, when checking the centrality metrics of videos connected to multiple topics in the network, we find that videos linked to all four topics have the highest betweenness centrality (0.017605), indicating their critical roles in **connecting various themes**. Similarly, the degree centrality (0.012048) and eigenvector centrality (0.086011) for these videos suggest a higher level of influence and connectivity within the network.

2.5 Sentiment Analysis

The following bar charts represent the VADER sentiment analysis (Figure 8):

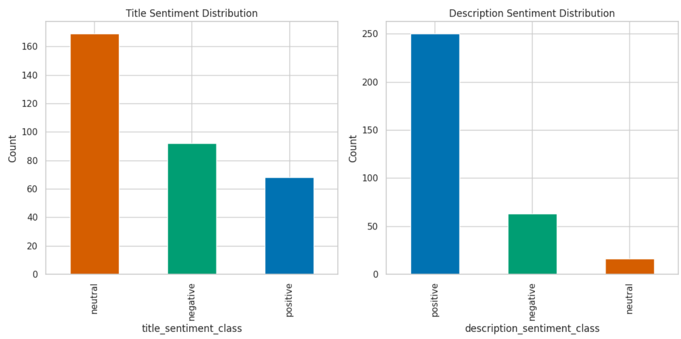

Figure 8. VADER Sentimental Classification of Title and Description

The following graphs represent RoBERTa's emotional detection analysis (Figure 9):

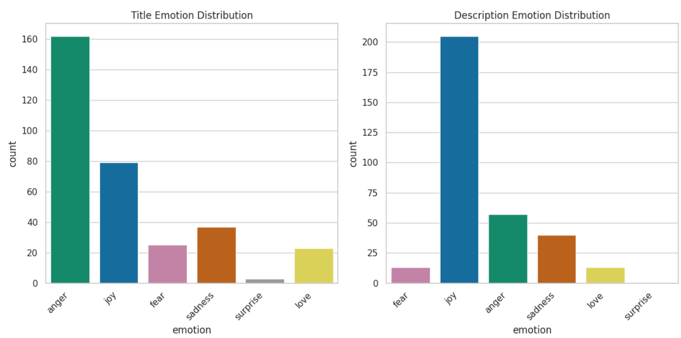

Figure 9. RoBERTa Sentimental Classification of Title and Description

VADER sentiment analysis: The majority of video titles are classified as neutral, with a significant portion falling under negative sentiment and fewer titles categorized as positive. In contrast, video descriptions show a dominant positive sentiment, with far fewer classified as negative and almost no neutral descriptions.

RoBERTa Emotion Detection: Anger is the most common emotion in titles, highlighting a possible tactic to **provoke** strong emotional reactions. Joy is the second most prevalent, but significantly less frequent than anger, showing a mix of emotionally charged and positive titles. As for descriptions, joy dominates aligning with the positive sentiment observed in the VADER analysis. Other emotions like anger, fear, and sadness are present but less common.

The sentiment analysis reveals that **neutral or negative titles**, paired with **positive descriptions**, likely serve to draw in viewers with curiosity or emotional hooks while maintaining engagement through a more optimistic tone (Figure 8). Emotion detection highlights the strategic use of anger in titles to provoke strong reactions, while joyful descriptions encourage positive viewer interaction, suggesting a deliberate shift in tone to maximize engagement (Figure Figure 9).

DISCUSSION AND LIMITATIONS

3.1 Discussion

In this study case of the famous YouTuber Rebal D, several primary findings have been put into light:

- Views, likes, and comments are interrelated metrics that correlate with one another.
- To be successful on YouTube a certain degree of consistency in video upload is required to not lose your fanbase.
- There are 4 specific topics that Rebal D covers in his content that evolve around video reactions, storytelling, and commentary videos. Overall, the network analysis shows **Topic 1's centrality**, emphasizing its role in engagement-related content.
- As showcased in the **sentiment analysis**, titles, often neutral or negative, seem designed to provoke **curiosity** or emotional responses, while more positive descriptions aim to build **engagement** and maintain viewer **interest**. This duality points to a tactic where initial emotional hooks are followed by a shift towards positivity in the body of the content, reinforcing the audience's desire to engage further.
- the **emotion analysis** highlights anger as the dominant title emotion, which aligns with the frequent use of **controversy or frustration** to capture attention, while joy dominates the descriptions, creating an emotional shift to maintain viewer satisfaction.

3.2 Limitations

This study, while providing insights into Rebal D's YouTube content strategies, has several limitations:

- **Case-Specific Focus:** The analysis is limited to a single YouTube channel, which may restrict the generalization of the findings to other creators or cases.
- **Sentiment Analysis** Accuracy is a bit low. We notice some emotional overlap. This could be due to Rebal D's heavy use of sarcasm that complicates sentiment detection or because Tools like VADER and RoBERTa may struggle with context and complexity. Further investigation and enhancement is required in this regard.
- **Data Scope:** The finite dataset may not capture all variations in video titles and descriptions, suggesting that larger datasets could strengthen future analyses.

CONCLUSION

This research on Rebal D's YouTube channel provides useful insights into how sentiment and emotive expressions are used in video content to engage viewers. We uncovered critical trends in content strategy underlining the complexities of sentiment interpretation when dealing with humor and sarcasm. While the findings show substantial advances in audience engagement tools, there are some constraints indicating that additional investigation is needed to explore broader applicability. Finally, this study adds to our understanding of digital content strategies while emphasizing the relevance of context in sentiment analysis.

AUTHOR

This work was done by **Nada Dallai**, Matricola 903644, as a part of the the course "Social Media and Web Analytics" taught by Professor **Fabiana Zollo** at the *University Ca' Foscari of Venice*.

ACKNOWLEDGMENTS This work was supported by the LaTeX template provided by the Brazilian Journal of Geophysics (BrJG).

REFERENCES

[1] Seong Soo CHA. The influence of youtube. *The Journal of Industrial Distribution & Business*, 15(9):1–9, 2024.

[2] Laurence Dessart and Valentina Pitardi. How stories generate consumer engagement: An exploratory study. *Journal of business research*, 104:183–195, 2019.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3: 993–1022, 2003. URL `http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf`.

[4] C. J. Hutto and E. E. Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the 8th International Conference on Weblogs and Social Media*, pages 216–225. Association for Computing Machinery, 2014.

[5] D. A. Aaker and D. M. Stayman. Measuring audience perceptions of the advertising budget. *Journal of Advertising Research*, 32(4):30–38, 1992.

[6] NetworkX. Networkx documentation, n.d. URL `https://networkx.org/documentation/stable/`.

[7] L. C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40(1):35–41, 1977. doi: 10.2307/3033543.

[8] S. P. Borgatti. Centrality and network flow. *Social Networks*, 27(1):55–71, 2005. doi: 10.1016/j.socnet.2004.11.008.

[9] P. Bonacich. Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology*, 2(1):113–120, 1972. doi: 10.1080/0022250X.1972.9989896.

[10] VADER. Vader sentiment analysis, n.d. URL `https://github.com/cjhutto/vaderSentiment`.

[11] Hugging Face Transformers. Roberta, n.d. URL `https://huggingface.co/docs/transformers/model_doc/roberta`.