Youtube Comment Analysis

Gabriel Lucchini, Nassim Lattab, Florian Posez, Mohamed Azzaoui

Professor : Themis Palpanas

Abstract—This report presents a YouTube comment analysis project aimed at evaluating the sentiment expressed in video comments. The main objective is to develop accurate sentiment analysis models to provide an overview of the overall sentiment of videos based on user comment analysis.

keywords—*YouTube Comments Analysis, Sentiment Analysis, Natural Language Processing, Data Mining, User Opinion, Text Analysis, Lexicon-based Sentiment Analysis*

Contents

1. Introduction

1.1. Problem Statement

YouTube, being one of the largest platforms for video content consumption, hosts a vast array of videos covering diverse topics and interests. With millions of users interacting through comments on these videos and the recent removal of the common dislike count on YouTube videos, evaluating user sentiment and feedback has become increasingly challenging. This change has obscured the visibility of negative sentiment expressed by viewers, making it difficult to gauge the overall impact and reception of videos. Consequently, there is a pressing need to develop robust methods for sentiment analysis of YouTube comments, compensating for the loss of the dislike count and providing a comprehensive understanding of user opinions and reactions to videos.

1.2. Importance

In today's digital landscape, quickly grasping the overall sentiment (positive or negative) surrounding a video is vital. Users form rapid impressions that influence their decision to engage with content. Therefore, having the ability to promptly gauge the general sentiment towards a video is invaluable. It aids content creators, marketers, and decision-makers in assessing impact, identifying issues, and adjusting strategies. Swift sentiment analysis is essential for content strategy, audience engagement, and brand reputation management.

1.3. Challenges

Despite its significance, sentiment analysis of YouTube comments presents several challenges. These include handling large volumes of comment data, processing the nuanced language used in comments, dealing with sarcasm, slang, and varying expressions of sentiment. Additionally, ensuring the accuracy and reliability of sentiment analysis models amidst the dynamic nature of user-generated content on YouTube poses another challenge. Overcoming these obstacles is essential to fully harness the potential of YouTube comments for understanding user opinions and guiding content and marketing decisions.

2. Related Work

2.1. Summary

We have selected several articles, all sharing a common theme: sentiment analysis or the management of large quantities of data from YouTube. These articles present various innovative approaches to tackle these subjects. Although their ultimate goals may differ from ours, they remain relevant for comparison, as they offer interesting perspectives on how to address these challenges. In the following subsection, we summarize the selected papers to provide insights into their methodologies and findings, enriching our understanding of the field.

2.2. Similar Papers and Approaches

Attention-based multimodal sentiment analysis and emotion recognition using deep neural networks [7]
The document introduces an innovative approach known as Lexicon and Syntax Enhanced Opinion Induction Tree (LSOIT) for emotion analysis. Various external knowledge sources, such as phrase structures and dependency relationships, are combined with machine learning techniques like graph neural networks. The goal is to enhance sentiment analysis accuracy by considering not only textual content but also sentence structure and semantics. Tests conducted on diverse datasets reveal that the LSOIT model outperforms existing methods in terms of performance. In summary, this method offers a novel and effective approach for studying emotions in text by leveraging different knowledge sources and advanced learning methods.

Sentiment analysis on youtube smartphone unboxing video reviews [11]
This study looks at sentiment analysis on YouTube videos featuring smartphone unboxings from Sri Lanka. It examines how consumer

purchasing decisions are increasingly influenced by online reviews, especially those that are in video format. The novel method, known as Lexicon and Syntax Enhanced Opinion Induction Tree (LSOIT), combines machine learning methods like graphical neural networks with external knowledge like phrase structure and dependency interactions. The findings demonstrate that the LSOIT model works better than the state-of-the-art techniques. In conclusion, this study offers a novel and successful approach to identifying emotions in texts by combining sophisticated learning techniques with a variety of knowledge sources.

Hate Speech Patterns in Social Media [12]
The book proposes a comprehensive approach to studying online hate speech, with a focus on the stigmatization of overweight individuals. It highlights the need to combine computational and qualitative methods to identify and interpret hate speech patterns, using techniques such as sentiment analysis, topic modeling, and discourse analysis. The various steps of the proposed methodology include data collection, data preprocessing, data analysis, and pattern identification. In summary, the text presents a thorough methodology for analyzing online hate speech and identifying associated linguistic patterns.

A review on sentiment analysis from social media platforms [10]
In this text, we delve into the study of sentiment analysis on social networks, highlighting its growing importance for businesses. It examines how social media analysis tools offer companies the ability to gather data about their market and customers, as well as effectively communicate commercial and marketing information. It references major players in the field, such as IBM, SAS, Microsoft, and SAP, along with the variety of tools available, ranging from social media management software to advanced analytics tools. Additionally, the text discusses the proliferation of patents related to sentiment analysis and the various applications of this analysis in different sectors such as marketing, politics, economics, health, and emergencies. Finally, it raises the issue of reproducibility and practical application of sentiment analysis methods, underscoring the diversity of approaches and the need for a better understanding of the advantages and limitations of each method.

Analyzing Social Media Data Using Sentiment Mining and Bigram Analysis for the Recommendation of YouTube Videos [8]
This paper combines sentiment analysis with graph theory to analyze user posts, likes/dislikes on various social media platforms to provide recommendations for YouTube videos. The focus is on the topic of climate change/global warming, which has sparked much alarm and controversy in recent years. The intention is to recommend informative YouTube videos to those seeking a balanced viewpoint on this subject and its key arguments/issues. To achieve this, the paper analyzes Twitter data, Reddit comments and posts, user comments, view statistics, and likes/dislikes of YouTube videos. The combination of sentiment analysis with raw statistics and linking users with their posts offers deeper insights into their needs and quest for quality information. Sentiment analysis provides insights into user likes and dislikes, while graph theory reveals linkage patterns and relationships between users, posts, and sentiment. We selected this document for our presentation as it closely aligns with our project.

Managing Diverse Sentiments at Large Scale [5]
This paper addresses the challenge of aggregating diverse sentiments at scale, particularly focusing on continuously updated data sources. It presents a theoretical framework for modeling sentiment diversity and introduces measures to capture this diversity from aggregated sentiment statistics. Robust and scalable indexing and storage methods are developed to handle diverse sentiments effectively. Additionally, an adaptive approach for identifying contradictions at different time scales is proposed. Experimental evaluations demonstrate the effectiveness of the proposed method in capturing contradictions, showcasing its superiority over relational databases in real-world scenarios.

Harvesting relational tables from lists on the web [1]
A novel technique for extracting tables from lists found on web pages is proposed. These lists often lack well-defined templates, making the extraction process challenging due to inconsistent delimiters and missing information. The technique operates in a fully unsupervised manner and is domain-independent. Initially, multiple sources of information are leveraged to split individual lines into multiple fields. Subsequently, these splits across multiple lines are compared to identify and rectify incorrect splits and alignments. Notably, a corpus of HTML tables, extracted from the web, is utilized to identify likely fields and optimal alignments. For each extracted table, an extraction score is computed reflecting confidence in its quality. Extensive experimental studies, involving both real web lists and lists derived from web tables, demonstrate the high accuracy of the technique in extracting tables. Moreover, the technique is applied to a large sample of approximately 100,000 lists crawled from the web, revealing the potential existence of tens of millions of useful and queryable relational tables extractable from web lists.

Survey on mining subjective data on the web [2]
This article provides an overview of the growing importance of Sentiment Analysis and Opinion Mining in Information Retrieval and Web data analysis. It highlights their significance in capturing sentiments and opinions at scale, particularly with the proliferation of user-generated content on platforms like blogs, wikis, and Web forums. The article discusses how opinion retrieval has become integral to search engines, enhancing user experiences by offering insights beyond traditional document retrieval. It emphasizes the value of aggregating opinions from product reviews for marketing purposes and understanding customer attitudes across various dimensions. Additionally, the article explores the evolution of Sentiment Analysis, Opinion Mining, and Contradiction Analysis as research fields, outlining recent advancements and future directions.

LSOIT: Lexicon and Syntax Enhanced Opinion Induction Tree for Aspect-based Sentiment Analysis [13]
This article proposes an innovative approach for aspect-based sentiment analysis (ABSA), a crucial task in sentiment classification. ABSA involves identifying sentiment aspect and polarity pairs within sentences. Unlike traditional methods that rely on serialization models, this approach utilizes graph neural networks enhanced with lexicon and syntax information. By integrating knowledge from lexicon (e.g., sememe knowledge) and syntax (e.g., phrase structures and dependency relationships), the proposed model, called Lexicon and Syntax Enhanced Opinion Induction Tree (LSOIT), aims to improve the accuracy of sentiment analysis. Experimental results on benchmark datasets demonstrate the superiority of LSOIT over state-of-the-art models, particularly in handling complex sentences and leveraging external knowledge effectively.

Analysis based on YouTube Channel Dataset [6]
Using machine learning analysis methods could provide an important reference basis for the company's precision marketing. This paper used a series of machine learning analysis methods on the YouTube channel dataset. These algorithms include linear regression, correlation analysis, k-means cluster analysis, and social network analysis. Based on the results of these algorithms, it can be concluded that a moderate negative association exists between length-view, length-ratings, and length-comments

where length is the time length for a specific video while views-ratings, views-comments, ratings-comments are positively associated.

Sentimental Analysis of Movie Tweet Reviews Using Machine Learning Algorithms [4]
This paper introduces a robust system architecture for sentiment analysis on microblogging platforms, particularly focusing on movie tweet reviews. The architecture incorporates various algorithms, including Multinomial Naive Bayes, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Bernoulli's Naive Bayes, and Random Forest, meticulously trained on annotated Twitter data. Thorough experimentation demonstrates the effectiveness of the methodology, which involves extensive data curation, intricate text preprocessing steps, and the utilization of diverse machine learning algorithms. The assessment includes metrics like accuracy, precision, recall, and F1-score, supported by techniques like cross-validation. The paper emphasizes the importance of preprocessing in improving sentiment classification accuracy and highlights the system's ability to extract valuable insights from microblog sentiments for informed decision-making.

Arabic Sentiment Analysis of YouTube Comments: NLP-Based Machine Learning Approaches for Content Evaluation [9]
This paper introduces an NLP-based model for classifying Arabic comments on YouTube as positive or negative. It aims to evaluate video quality without the need to watch it, especially in light of the recent removal of the dislike count feature on YouTube. The model, trained on a dataset of 4212 labeled comments, employs six classifiers including SVM, Naïve Bayes, Logistic Regression, KNN, Decision Tree, and Random Forest. Notably, the Naïve Bayes classifier demonstrates high accuracy. The study provides insights for content creators to improve content and audience engagement by analyzing viewer sentiments towards videos. Additionally, it fills a gap in the literature by offering a comprehensive approach to Arabic sentiment analysis, an area currently lacking exploration in the field.

2.3. Relation to Our Work

The examined articles present various approaches and methodologies related to our YouTube comment analysis project. They address aspects such as sentiment analysis, emotion content detection, and understanding user opinions from textual data. By exploring techniques such as social data analysis, advanced machine learning methods, and extraction of structured information from unstructured data, these articles provide valuable insights for our own YouTube comment analysis. Additionally, they highlight the importance of understanding user opinions in the current digital landscape, where user engagement and feedback greatly influence content perception and decisions of content creators, marketers, and policymakers.

Now, let's examine the relationship between our project and each of the papers presented:

1. **Attention-based Multimodal Sentiment Analysis and Emotion Recognition using Deep Neural Networks**: This paper introduces an innovative approach for emotion analysis.
2. **Sentiment Analysis on YouTube Smartphone Unboxing Video Reviews in Sri Lanka**: Explores sentiment analysis on YouTube videos.
3. **Hate Speech Patterns in Social Media**: Although the focus is on hate speech, this paper's methodology for analyzing patterns in social media aligns with our project's objective.
4. **A Review on Sentiment Analysis from Social Media Platforms**: Discusses sentiment analysis on social networks, providing insights into the importance of understanding user opinions.
5. **Analyzing Social Media Data Using Sentiment Mining and Bigram Analysis for the Recommendation of YouTube Videos**: Explores sentiment analysis on social media to recommend YouTube videos.
6. **Managing Diverse Sentiments at Large Scale**: Addresses the challenge of handling diverse sentiments.
7. **Harvesting Relational Tables from Lists on the Web**: Although focused on web data extraction, this paper highlights techniques for extracting structured information from unstructured data.
8. **Survey on Mining Subjective Data on the Web**: Discusses sentiment analysis in web data, emphasizing its importance across various domains.
9. **LSOIT: Lexicon and Syntax Enhanced Opinion Induction Tree for Aspect-based Sentiment Analysis**: Introduces an approach for sentiment analysis.
10. **Analysis based on YouTube Channel Dataset**: Utilizes machine learning analysis on YouTube data, offering insights into user behavior.
11. **Sentimental Analysis of Movie Tweet Reviews Using Machine Learning Algorithms**: Presents a methodology for sentiment analysis on microblogging platforms.
12. **Arabic Sentiment Analysis of YouTube Comments: NLP-Based Machine Learning Approaches for Content Evaluation**: Introduces a model for sentiment analysis on YouTube comments.

By examining these papers, we can draw upon their methodologies and insights to enhance our own YouTube comment analysis project, ultimately leading to a more comprehensive understanding of user sentiments and opinions.

3. Problem Formulation

3.1. Formal Definition

As previously mentioned, YouTube is filled with a wide variety of content from all over the world. The goal of this project was to gather infromations about the most popular videos on Youtube. Secondly to store all the collected data in an swift and easly accessible database. Then, to analyse their comments to extract an overall sentiment and opinions about the video. Given a dataset of YouTube comments, denoted as $D = \{c_1, c_2, ..., c_n\}$, where each comment c_i consists of text and metadata, our objective is to develop a robust sentiment analysis framework that accurately gauges the sentiment expressed in the comments.

Specifically, we aim to:

1. Classify each comment into one of three sentiment categories: positive, negative, or neutral.
2. Compensate for the absence of the dislike count on YouTube videos by accurately identifying and quantifying negative sentiment expressed by viewers.
3. Address challenges related to the dynamic nature of user-generated content, including large volumes of data, nuanced language, sarcasm, slang, and varying expressions of sentiment.
4. Ensure the accuracy and reliability of sentiment analysis models despite the dynamic nature of user-generated content on YouTube.

Our goal is to provide content creators, marketers, and decision-makers with a comprehensive understanding of user opinions and reactions to videos on YouTube, facilitating informed content strategy, audience engagement, and brand reputation management.

4. YouTube API and Data Mining

4.1. Solution description

Our solution involves using the YouTube Data API with Python. The API responses are presented as JSON formated data, which, while comprehensive, contains an abundance of useless information. Our primary objective in this project was to sift through these extensive

json responses to extract the most pertinent information, particularly focusing on the top 200 daily popular videos in both France and the United States. The goal was to gather informations on a daily basis to follow the top 200 videos overtime during a month. However, only the top 200 videos for each country are accessible, with popularity being the sole trend indicator. For each video retrieved, we further gather the 1000 most recent comments to ensure a comprehensive analysis of user sentiment. This step allows us to capture the latest feedback and opinions expressed by viewers, providing a real-time understanding of audience sentiment towards the videos.

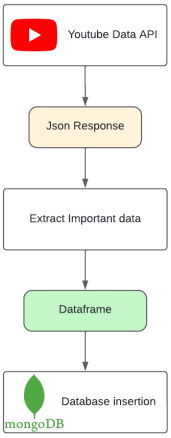

Figure 1. Data Mining Pipeline

Understanding the need for swift data access, we chose MongoDB, a structured database system known for its quick querying and data handling capabilities. MongoDB's flexible structure is ideally suited to manage the extensive data we gather, organizing it in a way that makes each data point easily accessible.

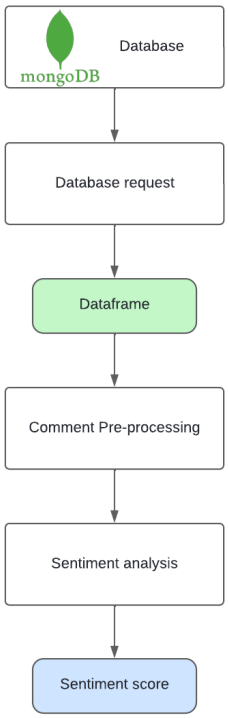

Figure 2. Comment Analysis Pipeline

4.2. Pseudo code

The Youtube Data API offers various request to access certain informations such as Channel, Videos, Comments. To handle all these possible request we created a class 'Request' to simplify the usage of recurent request (especially in while loops).

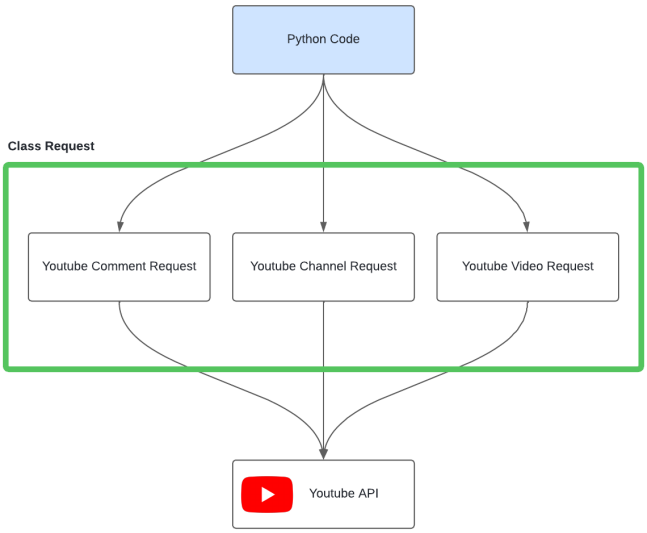

Figure 3. Class request

Since every Youtube API reponse is Json formatted we created multiple functions to select and gather important informations. Then these informations are formatted in dictionaries to be easly convertible into dataframes.

```
Function format_video_data(video_data):
    """ Structure raw video data """
    data = {
    "title": video_data.get('title'),
    "id": video_data.get('id'),
    "publishedAt":video_data.get('publishedAt'),
    "duration": video_data.get('duration'),
    "ViewCount": video_data.get('viewCount'),
    "likeCount": video_data.get('likeCount'),
    "commentCount": video_data.get('commentCount
    '),
    "tags": video_data.get('tags')
    }
    return DataFrame(data)

def get_video_data(youtubeAPI, video_Id):
    """ Request for most important video stats
    """
    request = Request(
        requestType=youtube.videos(),
        id=video_Id,
    )
    response = request.execute()

    rawData = response.get(items)
    return format_video_data(rawData)
```

Code 1. Fetching data functions

4.3. Data Management and Description

We chose MongoDB for our project due to its native compatibility with JSON, simplifying storage and manipulation of data from the YouTube API. With most of the retrieved data already in JSON format, using a database that supports it allows for more efficient work. Additionally, MongoDB offers high performance, even with large

data volumes, crucial for our project where we anticipate handling a significant amount of information from YouTube videos. Moreover, MongoDB is renowned for its user-friendly nature and intuitive query syntax, making database development and management simpler and quicker, especially in a NoSQL environment where data schemas can evolve over time.

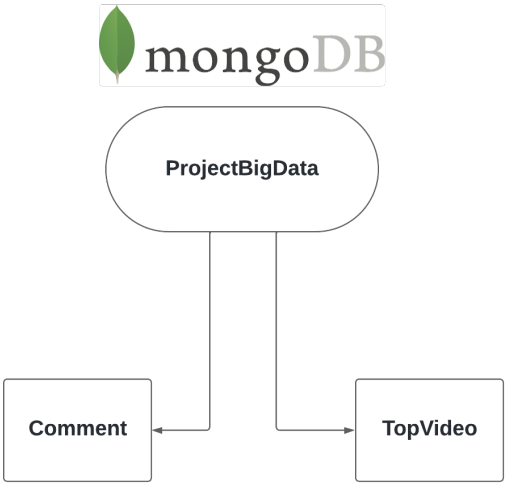

Figure 4. Comment Analysis Pipeline

In our project, we have decided to split our data into two distinct collections (a collection being a set of documents in MongoDB) as depicted in the figure above: one for comments and one for video information, in order to optimize our database architecture. This approach was chosen for several reasons. Firstly, it allows us to organize our data in a clear and efficient manner, facilitating management and retrieval. Secondly, by separating comments from video information, we can optimize performance by applying specific indexes and queries to each type of data. Additionally, this structure provides enhanced scalability, which is crucial given our projected data volume growth. Lastly, this division simplifies maintenance by enabling us to target operations such as backups, restores, and schema updates more precisely.

Below, two documents from each of the collections.


```
_id: ObjectId('6629314581477d6e37ae7f4b')
id : "UgxHYBQAE58AMk_Njg14AaABAg"
comment : "le grand retour tant attendu 😘"
likeCount : 4783
publishedAt : "2024-02-21T11:42:48Z"
updatedAt : "2024-02-21T11:42:48Z"
totalReplyCount : 52
videoID : "c1FLpmpo50c"
fetchedDate : "2024-03-04 13:15:00.801367"
preprocessed_comment : "grand retour tant attendu"
```


Figure 5. Document of collection Comment


```
_id: ObjectId('6629313b81477d6e37ae489b')
title : "Les vacances des Croûtons qui démarrent pas comme prévu... (c'est la c…"
id : "c1FLpmpo50c"
publishedAt : "2024-02-20T19:00:28Z"
duration : "PT29M28S"
ViewCount : 1209392
likeCount : 122489
commentCount : 3763
tags : "['Vacances des croûtons', 'croûton', 'team croûton', 'vlog', 'vacances…"
fetchedDate : "2024-02-21 18:55:50.000000"
topID : 1
region : "FR"
fetchedComments : true
Negative Comments : 80
Neutral Comments : 835
Positive Comments : 84
```


Figure 6. Document of collection Comment

As you can see, we have kept only the most relevant information in our database. Each piece of information will be used later for our various graphs.

5. Preprocessing and Sentiment Analysis (VADER)

5.1. Solution description

Once the YouTube comments are collected, we perform preprocessing to clean the data and prepare it for sentiment analysis. This preprocessing includes steps such as removing stop words, emoticons, and other non-essential elements from the text. The cleaned data is then ready for further analysis, including generating word clouds for statistical insights and conducting sentiment analysis using advanced techniques.

We employ state-of-the-art sentiment analysis techniques to extract sentiment polarity from YouTube comments. Our approach combines both lexicon-based methods and machine learning algorithms to accurately gauge sentiment. Specifically, we leverage the VADER lexicon-based tool for sentiment analysis due to its effectiveness in handling nuances of sentiment expressed in social media text.

VADER (Valence Aware Dictionary and sEntiment Reasoner) [3] is a lexicon and rule-based sentiment analysis tool specifically attuned to sentiments expressed in social media. It is fully open-sourced under the MIT License. For more details, please refer to the documentation, and additional information can be found on the corresponding GitHub page.

In our analysis, we utilize VADER to assess the sentiment of the comments and derive sentiment scores, categorizing them into 'Positive Comments', 'Negative Comments', and 'Neutral Comments'. This enables us to quantify the overall sentiment expressed towards a video, providing valuable insights into its reception among viewers.

5.2. Pseudo code

This pseudo code outlines the sentiment analysis process applied to the YouTube comments collected during the project.

The **preprocess_text** function is responsible for preprocessing the text data before sentiment analysis. It converts the text to lowercase, removes URLs, and eliminates non-alphanumeric characters. Then, it tokenizes the text, removes stopwords based on the language (English or French), and lemmatizes the remaining words to their base form. Finally, it joins the tokens back into a preprocessed text string.

The calculate **sentiment_counts** function calculates sentiment counts for a batch of comments. It initializes counters for positive, negative, and neutral sentiments. Then, for each comment, it uses the VADER sentiment analysis tool (sid) to obtain a sentiment score. If the compound sentiment score is greater than 0.05, the comment is classified as positive (incrementing the like count). If the score is less than -0.05, the comment is classified as negative (incrementing the dislike count). Otherwise, the comment is classified as neutral (incrementing the neutral count). Finally, it returns the counts for each sentiment category: like_count, dislike_count, and neutral_count.

These functions collectively preprocess the text data and calculate sentiment counts, providing valuable insights into the overall sentiment distribution among the YouTube comments analyzed in the project.

```
Function preprocess_text(text, language):
    If text is a string then
        convert_to_lowercase(text)
        remove_links(text,[http\S+])
        remove_special_characters(text,[^a-zA-Z0
    -9\s])

        tokens = tokenize(text)

        If language == 'US' then
            stop_words = get_stop_words('english
    ')
        Else If language == 'FR' then
            stop_words = get_stop_words('french'
    )
        Else
            return ''  # Unsupported language

        tokens = filter_stop_words(tokens,
    stop_words)
        lemmatizer = initialize_lemmatizer()
        tokens = lemmatize(tokens, lemmatizer)
        preprocessed_text = concatenate_tokens(
    tokens)
        return preprocessed_text
    Else
        return ''

Function calculate_sentiment_counts(comments):
    like_count = 0
    dislike_count = 0
    neutral_count = 0

    For each comment in comments do
        sentiment_score =
    calculate_sentiment_score(comment)

        If sentiment_score['compound'] > 0.05
    then
            like_count += 1
        Else If sentiment_score['compound'] <
    -0.05 then
            dislike_count += 1
        Else
            neutral_count += 1

    return like_count, dislike_count,
    neutral_count
```

Code 2. Preprocessing Text Implementation and sentiment analysis

5.3. Graphs and Results

In this section, we present a visual exploration of the data gathered from our sentiment analysis and content analysis of YouTube videos. Each figure provides valuable insights into the sentiment distribution, prevalent themes, and audience engagement patterns observed in both US and FR (France) videos. Let's delve into each figure to gain a comprehensive understanding of the findings:

The figure 7 illustrates the normalized sentiment ratios for US videos, sorted by the most positive sentiment. The graph reveals a predominant trend where most videos exhibit a higher proportion of positive sentiment compared to negative sentiment. Additionally, there is a notable presence of neutral sentiment across the videos, indicating a diverse spectrum of viewer opinions. This distribution underscores the generally positive reception of the analyzed US videos, with only a few outliers displaying a more balanced or negative sentiment ratio.

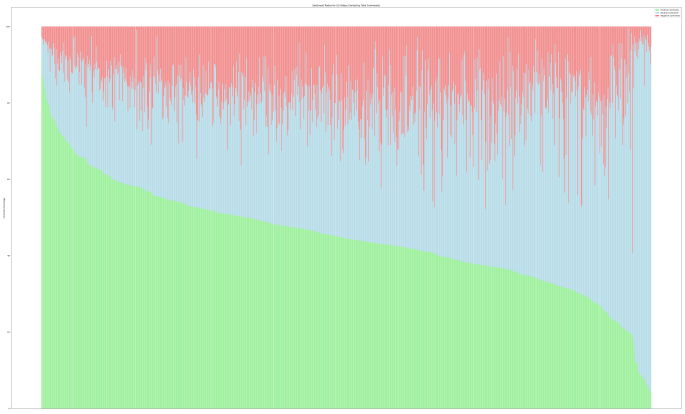

Figure 7. Normalized Sentiment Ratios for US Videos (Sorted by Most Positive)

In figure 8, the normalized sentiment ratios for US videos are sorted by the most negative sentiment. Contrary to the previous figure, most videos showcase a higher proportion of positive sentiment compared to negative sentiment. However, a significant portion of videos also exhibits a neutral sentiment, suggesting a mixed reception among viewers. While the overall sentiment leans towards the positive spectrum, the presence of more neutral sentiments highlights the complexity of audience reactions to the analyzed US videos.

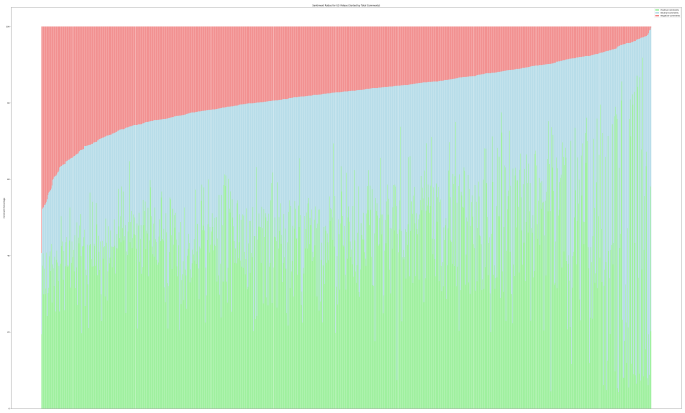

Figure 8. Normalized Sentiment Ratios for US Videos (Sorted by Most Negative

The figure 9 displays the normalized sentiment ratios for FR (France) videos, sorted by the most positive sentiment. Interestingly, a majority of videos exhibit a more neutral sentiment, with approximately an equal number of videos displaying more positive and negative sentiments. Despite this, there is a slight prevalence of positive sentiment among the analyzed FR videos. This distribution indicates a varied reception among viewers, with sentiments ranging from neutral to mildly positive.

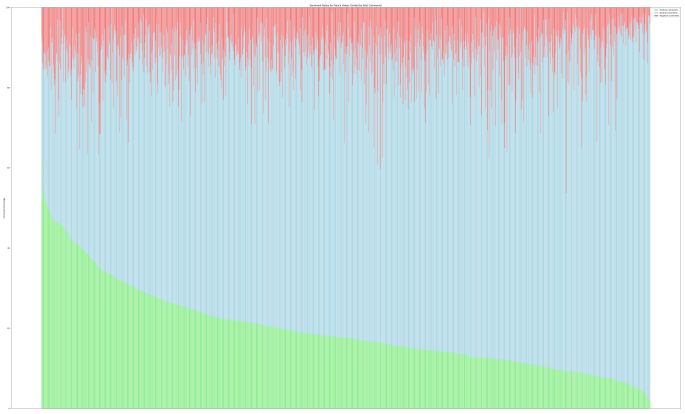

Figure 9. Normalized Sentiment Ratios for FR Videos (Sorted by Most Positive

In figure 10, the normalized sentiment ratios for FR videos are sorted by the most negative sentiment. Similar to the previous figure, a significant portion of videos demonstrates a neutral sentiment, with approximately equal numbers of videos showcasing more positive and negative sentiments. However, there is a slightly higher prevalence of positive sentiment compared to negative sentiment, suggesting a generally favorable reception among viewers despite some outliers with more negative sentiment.

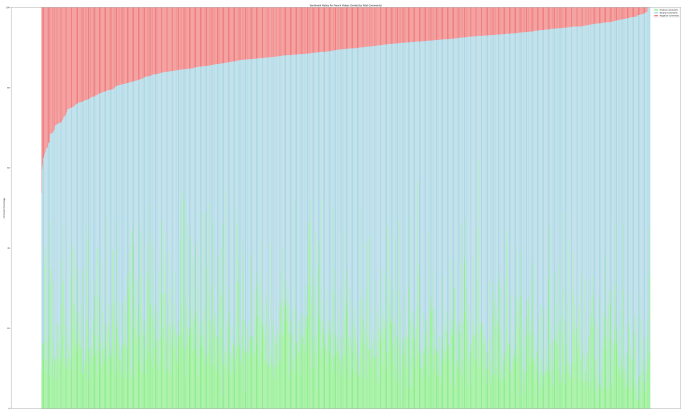

Figure 10. Normalized Sentiment Ratios for FR Videos (Sorted by Most Negative

Figure 11 presents a word cloud illustrating the most used tags in US videos. The word cloud prominently features terms such as 'new', 'game', and 'trailer', indicating common themes and topics among the analyzed US videos. These tags provide insights into popular content categories and subjects that resonate with the US audience, highlighting prevalent trends in video content on the platform.

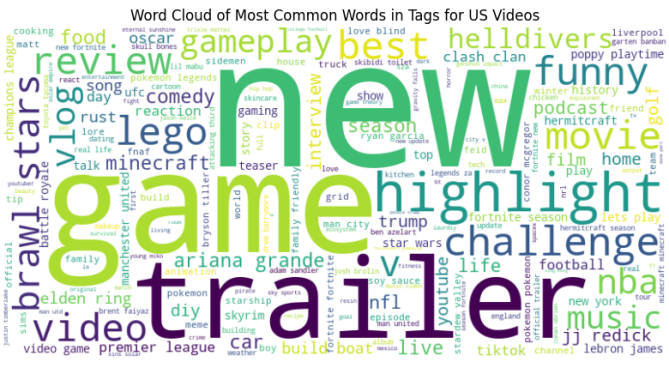

Figure 11. Word Cloud of most used US Tags

Figure 12 showcases a word cloud depicting the most used tags in FR (France) videos. Noteworthy terms such as 'sport', 'game', and 'live' dominate the word cloud, reflecting prevalent themes and interests among the analyzed FR videos. These tags offer valuable insights into the content preferences and audience engagement patterns within the French YouTube community, shedding light on the diverse range of topics and genres that attract viewers.

Figure 12. Word Cloud of most used FR Tags

Figure 13 presents a word cloud representing the most used words in US video titles. The word cloud highlights prominent terms such as 'new', 'v', 'music', 'official', and 'season', indicating common descriptors and keywords employed by content creators to attract viewers. These words offer insights into popular genres, formats, and content types prevalent in US video titles, providing valuable cues for understanding audience preferences and engagement patterns.

Figure 13. Word Cloud of most used words in US video titles

Lastly, Figure 14 showcases a word cloud displaying the most used words in FR video titles. Notable terms such as 'résumé', 'plus', 'V', and 'clip officiel' dominate the word cloud, reflecting common phrases and descriptors utilized in French video titles. These words offer insights into popular content themes, formats, and genres that resonate with the French YouTube audience, providing valuable cues for content creators and marketers seeking to engage with this demographic.

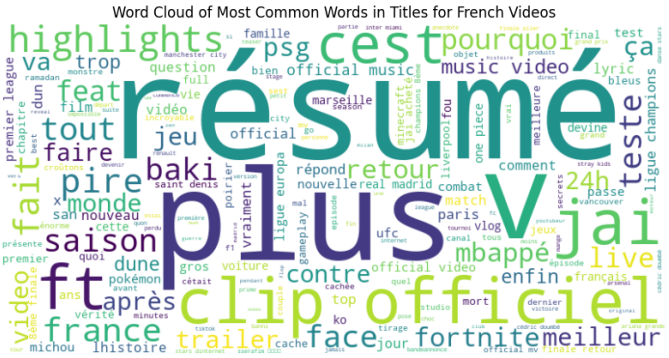

Figure 14. Word Cloud of most used words in FR video titles

Additionally, it's important to note that the frequent appearance of 'V' in the 2 word cloud of most used words in video title corresponds to the name of a popular K-pop group that released a highly popular song during the time of data collection, contributing to its prevalence in the analyzed video titles.

6. Conclusion

In conclusion, our YouTube comment analysis project aimed to provide valuable insights into user sentiments and opinions expressed in video comments, leveraging advanced sentiment analysis techniques. Through the comprehensive analysis of YouTube comments from top videos in both the United States and France, we gained valuable insights into audience engagement patterns, prevalent themes, and sentiment distributions.

Our findings revealed notable trends in the sentiment expressed by viewers towards videos from both regions. In the US, most videos exhibited a predominantly positive sentiment, with a notable presence of neutral sentiments across the analyzed videos. Conversely, the sentiment distribution in French videos showcased a more varied reception, with a majority of videos demonstrating a neutral sentiment and an almost equal number of videos displaying positive and negative sentiments.

Moving forward, the insights and data obtained from our YouTube comment analysis project offer numerous opportunities for future work and research endeavors:

- **Content Recommendation Systems**: Leveraging the insights gained from sentiment analysis, future work could focus on developing more personalized and targeted content recommendation systems. By understanding viewer preferences and sentiments, recommendation algorithms can suggest relevant content that aligns with users' interests, thereby enhancing user satisfaction and engagement.
- **User Engagement Strategies**: Utilizing sentiment analysis data, marketers and content creators can develop tailored user engagement strategies to foster meaningful interactions and build stronger connections with their audience. By understanding the sentiments and preferences of their viewers, content creators can optimize content delivery, engagement tactics, and community management efforts.
- **Predictive Analytics**: By analyzing historical sentiment data and user engagement patterns, predictive analytics models could be developed to forecast future trends in audience sentiment and behavior. These predictive insights can empower content creators and marketers to proactively adapt their strategies and content offerings to meet evolving audience preferences and maximize engagement.

Overall, our analysis contributes to a better understanding of user sentiments and preferences on YouTube, empowering content creators, marketers, and decision-makers to make informed decisions regarding content strategy, audience engagement, and brand reputation management.

References

[1] H. Elmeleegy, J. Madhavan, and A. Halevy, "Harvesting relational tables from lists on the web", *The VLDB Journal*, vol. 20, no. 2, pp. 209–226, 2011.

[2] M. Tsytsarau and T. Palpanas, "Survey on mining subjective data on the web", *Data Mining and Knowledge Discovery*, vol. 24, no. 2, pp. 478–514, 2012.

[3] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text", *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8, no. 1, pp. 216–225, May 2014.

[4] A. Amolik, N. Jivane, M. Bhandari, and M. Venkatesan, "Twitter sentiment analysis of movie reviews using machine learning techniques.", *International Journal of Engineering and Technology*, vol. 7, pp. 2038–2044, Jan. 2016.

[5] M. Tsytsarau and T. Palpanas, "Managing diverse sentiments at large scale", *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 11, pp. 3028–3040, 2016.

[6] X. Zhang and Z. Sun, "The statistical analysis based on youtube channel dataset", in *Proceedings of the 2021 1st International Conference on Control and Intelligent Robotics*, Association for Computing Machinery, 2021, pp. 376–382.

[7] A. Aslam, A. B. Sargano, and Z. Habib, "Attention-based multimodal sentiment analysis and emotion recognition using deep neural networks", *Applied Soft Computing*, vol. 144, p. 110 494, 2023.

[8] K. McGarry, "Analyzing social media data using sentiment mining and bigram analysis for the recommendation of youtube videos", *Information*, vol. 14, no. 7, p. 408, 2023.

[9] D. A. Musleh, I. Alkhwaja, A. Alkhwaja, *et al.*, "Arabic sentiment analysis of youtube comments: Nlp-based machine learning approaches for content evaluation", *Big Data and Cognitive Computing*, vol. 7, no. 3, 2023.

[10] M. Rodríguez-Ibánez, A. Casánez-Ventura, F. Castejón-Mateos, and P.-M. Cuenca-Jiménez, "A review on sentiment analysis from social media platforms", *Expert Systems with Applications*, vol. 223, p. 119 862, 2023.

[11] S. Sally, "Sentiment analysis on youtube smart phone unboxing video reviews in sri lanka", 2023.

[12] V. U. Wanniarachchi, C. Scogings, T. Susnjak, and A. Mathrani, "Hate speech patterns in social media: A methodological framework and fat stigma investigation incorporating sentiment analysis, topic modelling and discourse analysis", *Australasian Journal of Information Systems*, vol. 27, 2023.

[13] H. Wu, D. Zhou, C. Sun, Z. Zhang, Y. Ding, and Y. Chen, "Lsoit: Lexicon and syntax enhanced opinion induction tree for aspect-based sentiment analysis", *Expert Systems with Applications*, vol. 235, p. 121 137, 2024.