International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 12 Issue: 1
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 December 2023*

Exploring Sentiment Analysis in Social Media: A Natural Language Processing Case Study

Manish Rana
Department OF Computer Engineering,
Thakur College of Enginering And Technology
Mumbai, India
manish.rana@thakureducation.org,manishrana23@gmail.com

Rahul Khokale
Department of Computer Science and Engineering,
G H Raisoni University,
Saikheda , India
softrahul@mail.com

Sunny Sall
Department of Computer Engineering,
St. John College of Engineering and Management ,
Palghar , India
sunny_sall@yahoo.co.in

Abstract—Social media plays an integral role in our daily lives, influencing and reflecting global perspectives through the consumption and creation of content. Platforms like YouTube are incredibly active, with a constant influx of video uploads, views, and comments. While the YouTube app allows us to browse videos and comments, it offers only a limited glimpse into the interests and trends of others. Analysing this vast data pool, encompassing diverse language styles, presents a significant challenge. This article delves into the YouTube Data API and its application in Python for accessing raw data. The process involves data cleaning using advanced Natural Language Processing (NLP) techniques, harnessing Python-based machine learning to explore social media interactions, and automating the extraction of trends and influential factors. The journey towards trend analysis is meticulously detailed, featuring examples that leverage a variety of open-source Python tools.

Keywords- NLP; Natural Language Processing; social media data; YouTube; Named entity recognition; NER; Key-phrase extraction.

I. INTRODUCTION

In the digital realm, social media is a bustling global arena, teeming with activity around the clock. However, handling text data from social media presents a unique challenge due to the absence of established conventions and guidelines. This data, stemming from diverse geographical locations, languages, writing styles, and topics, defies easy standardization when it comes to processes like data cleansing, extraction, and the application of Natural Language Processing (NLP) techniques.

To access social media data, we rely on official and accessible APIs. Notable examples include the YouTube Data API and the Twitter API. It's essential to align your specific use case with the guidelines of the chosen API. In this endeavor, we'll focus on the YouTube Data API, addressing common challenges and providing efficient tools for data retrieval.

NLP tasks become feasible when there's accessible text data for analysis. We'll delve into the intricacies of textual anomalies within social media content. Furthermore, we'll thoroughly examine the process of cleansing noisy text using Python methods and open-source resources.

Social media data is rich with information, encompassing factors like content popularity, likes, dislikes, and various metrics. We'll explore the analysis of both statistical data and text data obtained from the YouTube API, evaluating their significance. To conclude, we'll perform trend analysis using open-source Python tools, leveraging social media data, statistics, NLP techniques for data refinement, and named entity recognition (NER) to craft a comprehensive analytical narrative.

.

II. NATURAL LANGUAGE PROCESSING

Natural Language Processing (NLP) [1] is the art of computer systems interacting with human language, encompassing both written and spoken forms. This interdisciplinary field, as outlined in [Wik21], merges linguistics, computer science, and artificial intelligence. NLP's primary focus revolves around the synergy between computers and human language, with a particular emphasis on enabling computers to efficiently process and analyze extensive volumes of natural language data. The ultimate goal is to equip computers with the capacity to "comprehend" the content of documents, delving into the intricate nuances of the language within. Some of the common NLP tasks on text data include the following.

1) Named-entity recognition

Named Entity Recognition (NER),[2] also referred to as Named Entity Identification, Entity Chunking, or Entity Extraction, plays a critical role in information extraction. Its primary objective is to detect and categorize named entities within unstructured text. These entities are classified into predefined categories, including personal names, organizations, locations, medical codes, time references, quantities, monetary values, percentages, and more. To address NER tasks encompassing a wide array of entity types, Python offers robust libraries such as SpaCy [HMVLB20] and NLTK [BKL09].

2) Key-phrase extraction

Key-phrase extraction is an automated process that aims to identify a concise set of phrases that effectively summarize the content of a given unrestricted text document, as described in [BSMH+18]. Notable resources for performing key-phrase extraction, as mentioned in this source, include Gensim [RS11] and RAKE-NLTK[#]_. Another avenue for key-phrase extraction involves harnessing NLTK [BKL09] techniques, a feature integrated into the pyYouTubeAnalysis [Sin21] library.

3) Unigrams/Bigrams/Trigrams analysis

Breaking down text into individual words, pairs of consecutive words, or sequences of three consecutive words, and scrutinizing the patterns of their occurrences. This approach unveils valuable insights into the structure and context of the language used in the text.

4) Custom classifier building (public dataset -> features -> ML models)

If off-the-shelf solutions aren't available for your NLP task, crafting custom models using accessible data, NLP libraries like NLTK2, SpaCy3, and Gensim4, along with machine learning libraries like scikit-learn5, is a viable option.

5) Others

Tokenization, part-of-speech tagging, lemmatization, stemming, word sense disambiguation, topic modelling, sentiment analysis, and text summarization are some widely used NLP tasks. This list is not exhaustive.

Machines can process a significantly larger volume of text samples than humans can in a day. Leveraging machines for NLP tasks, along with various Python processing solutions like multiprocessing, enables the analysis of vast amounts of data within a reasonable timeframe.

Potential use cases include the following.

1) Analytics, intelligence, and trends

Analysing text patterns involves assessing word frequencies, language nuances, and merging textual data with other relevant information, encompassing topics, sentiment analysis, NLP model outputs, or various combinations thereof.

2) Story telling

Analysing text through various NLP techniques, combined with statistical and available data, transforms raw information into an insightful narrative. This narrative unveils hidden patterns within the data, fostering a deeper understanding. Depending on the dataset, a time-window analysis can track evolving patterns over time, encompassing word usage, topics, text length, or various combinations thereof.

https://towardsdatascience.com/extracting-keyphrases-from-text-rakeand-gensim-in-python-eefd0fad582f
https://pypi.org/project/rake-nltk/
https://scikit-learn.org/
https://www.nltk.org/
https://spacy.io/
https://radimrehurek.com/gensim/
https://scikit-learn.org/
https://docs.python.org/3/library/multiprocessing.html

III. Social Media APIs

Numerous social media platforms offer APIs for programmatically accessing publicly available data and your own published data. Regardless of your data utilization intentions, it's crucial to adhere to the API's guidelines and terms of service.

YouTube provides [3] various types of requests, including search, video, channel, and comments, which can be accessed through the YouTube Data API. You can find comprehensive information and guidance in the YouTube Data API documentation7, including steps to get started8, such as registering a project, enabling it, and using the generated API key. With this key, you can commence making requests to the API for data retrieval.

A. Gotchas

However, there are some important considerations to keep in mind while using the YouTube Data API. Here are a few noteworthy points to be aware of:

***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 December 2023*

1) *Rate limits*

The API key you've registered comes with a daily usage limit. The rate at which you consume this limit depends on the type of requests you make. It's worth noting that the API won't warn you in the response if you're approaching your daily limit, but it will throw an error once that limit is exceeded. To prevent unexpected behaviour and premature script termination, it's essential to understand how your application will respond when you reach the quota.

2) *Error handling*

When attempting to query a private video, comment, or channel, the API [4] will generate an error. This can potentially disrupt your code, particularly if you're making multiple requests in a loop. Implementing effective error handling can enhance the automation of your process and ensure smoother handling of these expected errors.

B. *Interacting with the YouTube*

Data API offers several avenues:

Some of them are as follows.

1) Utilize the "Try this API" section in the API web explorer9.
2) Construct you code by referencing the API Documentation examples10.
3) Employ Open-source tools for streamlined API interaction

I. YouTube Data API Wrappers11: These are libraries that serve as wrappers, offering a convenient means to access and utilize the YouTube Data API V3.
II. You Tube Analysis: pyYouTubeAnalysis: This library empowers users to conduct searches, gather videos and comments, and customize search parameters such as keywords, timeframe, and type. Notably, it incorporates robust error handling to ensure uninterrupted code execution when encountering unforeseen issues while interacting with the pyYouTubeAnalysis Data API. Furthermore, YouTube Analysis offers additional features, including NLP methods for preprocessing social media text, as discussed in the subsequent Data Cleaning Techniques section, which utilizes NLTK based Key phrase extraction

https://developers.google.com/youtube/v3/docs

https://developers.google.com/youtube/v3/getting-started

and SpaCy based Named Entity Recognition (NER) that runs entity extraction within the text.

IV. Social Media / YouTube Data Noise

Text fields on YouTube are present in various locations, including video titles, descriptions, tags, comments, channel titles, and channel descriptions. Video titles, descriptions, tags, and channel information are typically provided by the content or channel owner. In contrast, comments are generated by individual users in response to videos, using their own words and language.

Challenges within this data source stem from the diversity of writing styles, languages, and topics. Figure 1 illustrates examples of language diversity. On social media, people frequently employ non-traditional abbreviations and may use less common abbreviations. In addition to non-standard abbreviation usage, you'll encounter different languages, variable text lengths, and the use of emojis by commenters.

A. *Data Cleaning Techniques*

To address the noise often found on YouTube and other social media platforms, the following data cleaning techniques have proven to be effective:

1) Removing URLs

Social media text data [5] is often cluttered with URLs. Depending on the specific task, removing URLs can be a valuable step in cleaning the text. Eliminating URLs before applying Keyphrase or NER extractions has been shown to yield cleaner results. This implementation is also integrated into pyYouTubeAnalysis. import re

```
URL_PATTERN = re.compile (
    r"https?://\S+|www\.\S+", re.X
)

def remove_urls(txt):
    """
    Remove urls from input text
    """ clean_txt = URL_PATTERN.sub(" ", txt)
    return                          clean_txt
```

2) Removing emojis

Emojis are extensively utilized on social media to convey emotions. Emojis can offer advantages in certain NLP tasks [6], particularly in sentiment analysis that relies on emoji-based detection. However, in many other NLP tasks, the removal of emojis from text serves as a valuable cleaning method, enhancing the quality of the processed results. In the context of named-entity recognition and Keyphrase extraction, specific emojis are occasionally misidentified as locations or nouns (NN or NNP), which can adversely affect NLP performance.

https://developers.google.com/youtube/v3/docs/search/list

***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 December 2023*

https://developers.google.com/youtube/v3/quickstart/python
https://github.com/rohitkhatri/youtube-python,
https://github.com/snssdks/python-youtube
https://github.com/jsingh811/pyYouTubeAnalysis

Eliminating emojis before subjecting the text to named-entity recognition or Keyphrase extraction has proven to produce more accurate outcomes. This implementation is also integrated into pyYouTubeAnalysis.

```
import re
EMOJI_PATTERN = re.compile( "[\U00010000-
      \U0010ffff]", flags=re.UNICODE
)

def remove_emojis(txt):
      """
      Remove emojis from input text
      """ clean_txt = EMOJI_PATTERN.sub(" ", txt)
      return                          clean_txt
```

3) Spelling / Typos Corrections

Some NLP models can excel with a particular style of language, but on social media, you often encounter misspelled words or typos. Open-source tools like PySpellChecker [LT16]13, Autocorrect14 and Textblob [Lor18] are available for performing spelling and typo corrections.

4) Language detection and translations

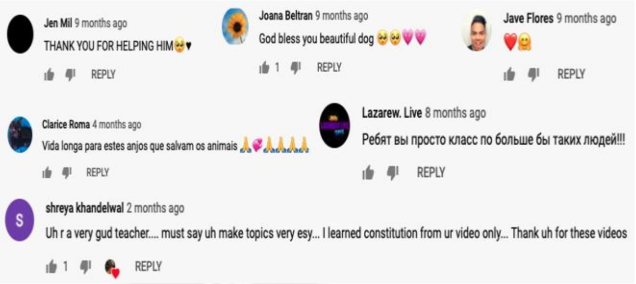
Creating NLP solutions for different languages [7] can be challenging yet rewarding. If you've developed NLP methods for English and encounter foreign languages, language detection and translation to English can be a valuable solution, albeit with its own set of challenges like detection and translation quality. Python libraries such as langdetect [Shu10], Pycld215, Textblob [Lor18], and Googletrans16 can be utilized for language detection. For language translations, options like Translate17 and Googletrans are available.

V. Trend Analysis Case Study

In 2020, the COVID-19 pandemic had a profound impact on the world, prompting swift changes to curb the virus's spread. One sector significantly affected was the travel and hospitality industry. Figure 218 illustrates the fluctuations in flight search trends in the US, both domestically and internationally, from February to November 2020, using Kayak data. Just before the implementation of lockdowns and restrictions in March, a notable surge in flight searches is evident, reflecting the efforts of people to return home in response to impending travel limitations.

https://pypi.org/project/pyspellchecker/
https://pypi.org/project/autocorrect/
https://pypi.org/project/pycld2/
https://pypi.org/project/googletrans/
https://pypi.org/project/translate/

A. YouTube comments representing writing style diversity.



Jen Mil 9 months ago
THANK YOU FOR HELPING HIM🥺♥
REPLY

Joana Beltran 9 months ago
God bless you beautiful dog 🥺🥺💗💗
1 REPLY

Jave Flores 9 months ago
❤️🤗
REPLY

Clarice Roma 4 months ago
Vida longa para estes anjos que salvam os animais 🙏💞🙏🙏🙏🙏
REPLY

Lazarew. Live 8 months ago
Ребят вы просто класс по больше бы таких людей!!!
REPLY

shreya khandelwal 2 months ago
Uh r a very gud teacher.... must say uh make topics very esy... I learned constitution from ur video only... Thank uh for these videos
1 REPLY

Figure. 1: Random sample of YouTube comments representing writing style diversity.

B. Domestic and international flight

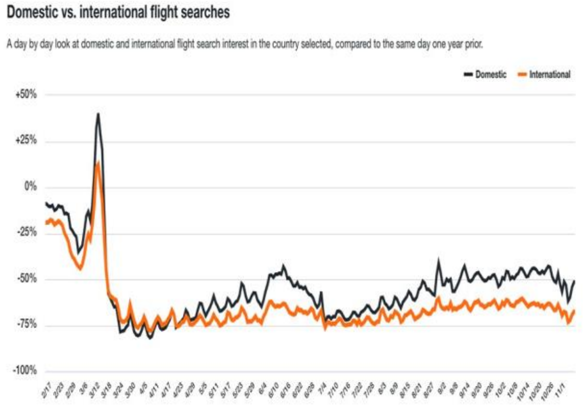


Figure . 2: Domestic and international flight search patterns in 2020.

C. Global flight search pattern

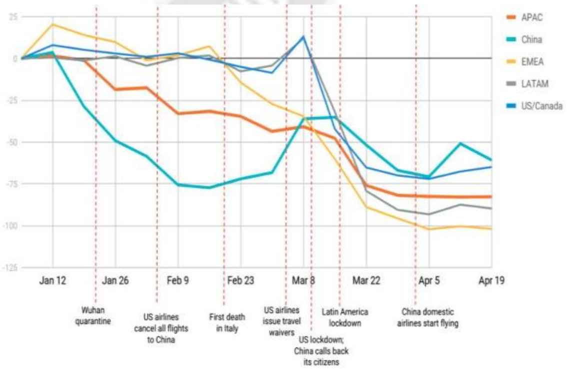


Figure 3: Global flight search patterns in 2020.

Figure 3 provides a stark visualization of the significant decline in flight searches, signifying the global impact of COVID-19. The timeline from January 2020 for China and March 2020 for most other regions witnessed the most substantial effects as travel was curtailed due to COVID-related events and restrictions.

This reduction in flight searches was mirrored by decreased hotel searches, as shown in Figure 4.

To gain a deeper understanding of the content consumption patterns during these periods on YouTube, a search was conducted using the pyYouTubeAnalysis library to gather videos related to "travel vlogs." Travel vlogs are a popular genre on YouTube, offering reviews, advice, and glimpses of various destinations, inspiring travel plans. These videos typically feature individuals documenting their journeys to different locations.

Statistically, as seen in Figures 5, 6, and 7, travel vlogs exhibited growing interest, aligning with the rise in online content consumption up to 2019. However, in 2020, there was a notable decline in average views, comments, and likes on travel vlog videos, with views dropping by 50% compared to the previous year.

To delve deeper into the differences [8] between travel vlog content consumed in 2019 and 2020, a monthly data analysis was conducted. Figures 8, 9, and 10 present a month-over-month comparison between the two years, examining audience engagement patterns. These trends reflect the decline starting in March when the impact of COVID-19 became widespread. Figure 11a & 11b further illustrates the engagement shift between 2019 and 2020. The trend initially rose but began declining as many regions implemented stay-at-home orders and lockdowns. The upward trend reemerged in July, coinciding with the easing of travel restrictions in Europe. The engagement chart for "travel vlog" content closely corresponds with the flight search trend shown in Figure 2.

Nevertheless, people continued to create travel vlogs and engage with such content between June and September 2020, despite reduced travel opportunities. This prompts questions about the nature of these videos, their content, creators, and the discussions in the comments.

Figure 12 presents a word cloud generated through Keyphrase extraction using the pyYouTubeAnalysis implementation, which involves data cleaning techniques before Keyphrase extraction. These techniques helped eliminate noise and enhance result quality. Word cloud visualizations represent term occurrences, with the size of a term in the cloud proportional to its frequency. Notably, in 2020, travel content emphasizing activities conducive to social distancing, such as hiking, beach trips, and road travel, gained prominence. Location names such as Italy, France, and Spain also appeared frequently in the videos, reflecting viewers' interests and aspirations.

https://www.kayak.com/news/category/travel-trends/

https://www.sojern.com/blog/covid-19-insights-on-travel-impact-hotelagency/

https://www.sojern.com/blog/covid-19-insights-on-travel-impact-hotelagency/

https://pypi.org/project/wordcloud/,
https://www.wordclouds.com/

We've explored[10] the content that garnered the most engagement, now let's delve into the creators behind this engaging content during the summer and fall of 2020, using engagement statistics and video analysis. The YouTube influencer channels that drove high engagement during this period include:

1) 4K Walk – A YouTube channel specializing in walking tours across Europe and America.
2) BeachTuber – A YouTube channel featuring vlogs from various European beaches.
3) Beach Walk – A YouTube channel highlighting different beaches across Europe and America.
4) DesiGirl Traveller – A YouTube channel creating videos about travel in India.
5) Euro Trotter – A YouTube channel dedicated to travel in Europe.

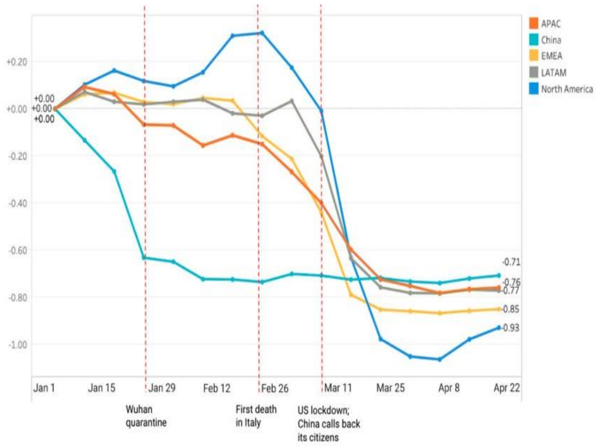
A few examples of comments that were being left by audiences of such videos are as follows.

> Audiences left various comments on these videos, and some examples include:
> "I'm going to Sorrento in 10 days, and I'm so excited. I've been watching tons of Sorrento and Italy vlogs, and yours are so lush X) <3"
> "Did they require you to have a prior COVID test?"
> "I loved the tour; it looked like you guys had fun. I'm going there next week. How long ago were you there, and were there lots of restrictions and closures due to COVID?"
> "Great video, man. This place looks amazing. I have never been to Iceland; I would love to visit someday. Honestly, I can't wait for the lockdown to be lifted so I can start traveling again. Thanks for sharing your experience. :)".

These comments often revolved around inquiries regarding travel bans due to COVID, pre-travel COVID test requirements, and the anticipation of traveling again. Commenters frequently mentioned specific locations, which were extracted using named-entity recognition (NER) with the pyYouTubeAnalysis library. The process involved passing comments through URL and emoji removal before location extraction, resulting in cleaner results and reduced manual filtering. Figure 13 displays a word cloud of the most frequently mentioned locations in the comments, featuring European, Asian, and American destinations. These locations align with the easing of travel restrictions in various places.

This analysis, along with data collection from social media, keyphrase extraction, and NER, was conducted using the pyYouTubeAnalysis library [Sin21]27. Similar analyses for content beyond "travel vlogs" can be performed for custom time windows using similar tools and the NLP libraries mentioned in this endeavour.

***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 December 2023*



Source: Sojern Data, YoY change in Hotel searches through April 26, 2020

Figure. 4: Hotel booking search patterns in 2020.

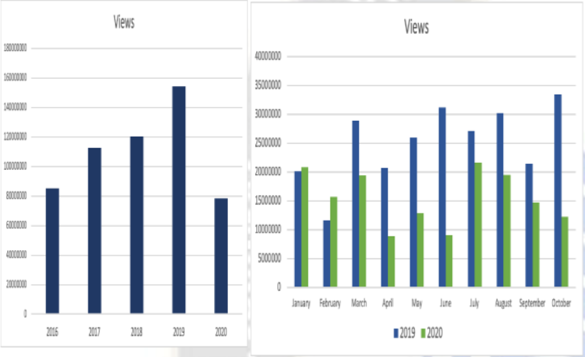


Figure. 5 , 6 & 7: Yearly video views.

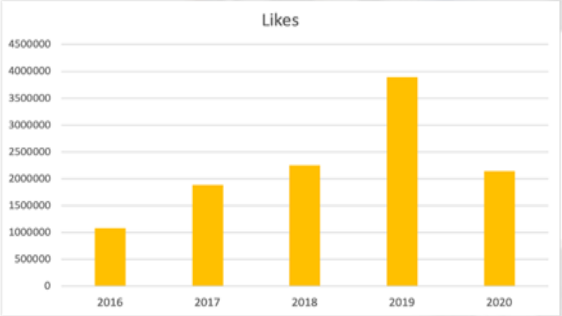


Figure. 8: Monthly video views for 2019 and 2020.

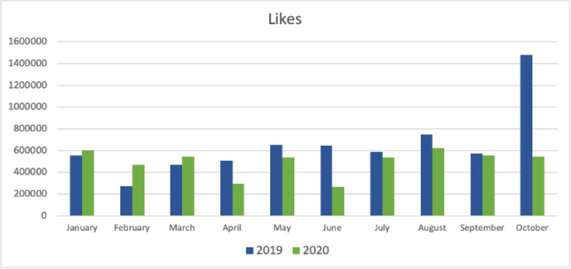


Figure. 9: Monthly video likes for 2019 and 2020.

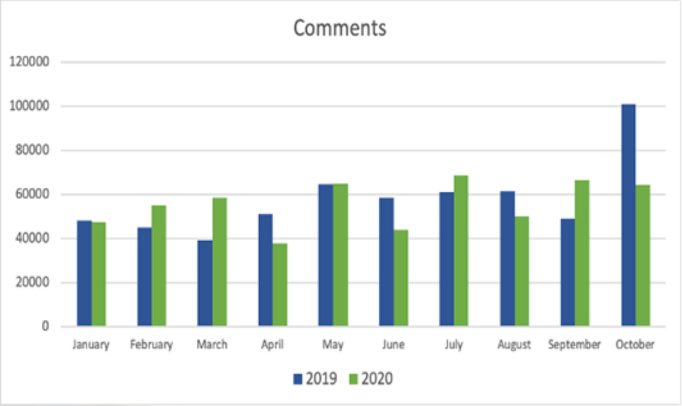


Figure. 10: Monthly video comments for 2019 and 2020

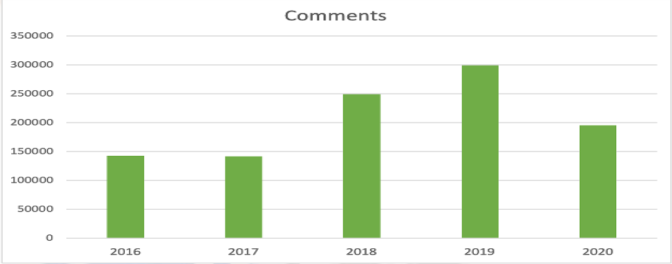


Figure.11a: illustrates the engagement shift between 2019 and 2020

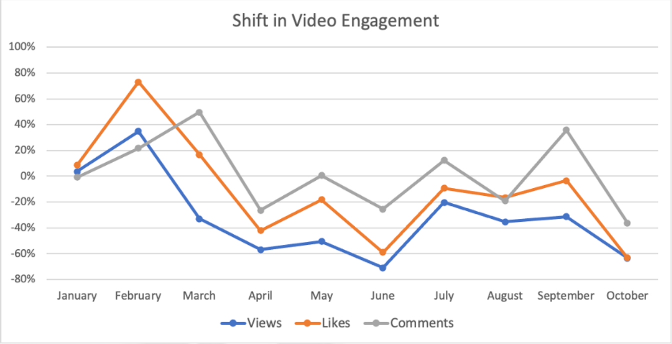


Figure.11b Shift of Vedio engagement shift between 2019 and 2020

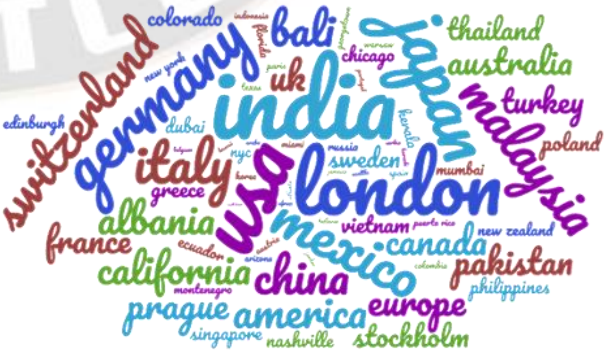


Figure 12 presents a word cloud generated through Keyphrase extraction

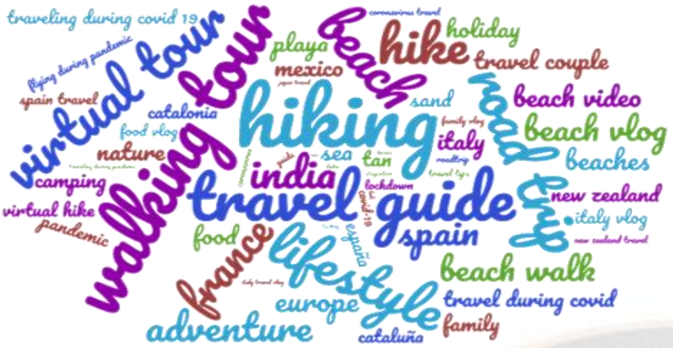


Figure 13 Word cloud of location names used in comments.

VI. Conclusion

User content creations and interactions via text on social media platforms contain mixed writing styles, topics, languages, typing

Variations in content and language, including errors, freeform emojis, and abbreviations, present challenges when performing NLP tasks on social media data. Cleaning techniques like emoji removal, hyperlink removal, language detection and translations, and typo corrections have proven valuable for preparing and pre-processing such text. Subjecting the text to these methods before applying other Natural Language Processing (NLP) techniques, such as Keyphrase extraction and named-entity recognition, results in cleaner output.

Social media data is rich in both textual content and statistical information, capturing human engagement and content preferences. Integrating these statistics with insights derived from NLP techniques like named-entity recognition (NER) and Keyphrase extraction proves valuable for trend analysis, analytics, and uncovering correlations and user engagement affinities in the social media landscape.

[24] https://youtube.com/c/4KWALK
[25] https://youtube.com/c/BeachTuber
[26] https://youtube.com/c/BeachWalk
[27] https://youtube.com/c/DesiGirlTraveller
[28] https://youtube.com/c/EuroTrotter
[29] https://github.com/jsingh811/pyYouTubeAnalysis

Acknowledgment

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression, “One of us (R.B.G.) thanks . . .” Instead, try “R.B.G. thanks”. Put applicable sponsor acknowledgments here; DO NOT place them on the first page of your paper or as a footnote.

References

[1] Steven Bird, Ewan Klein, and Edward Loper. Natural Language Processing with Python. 01 2019.

[2] Kamil Bennani-Smires, Claudiu Musat, Andreea Hossmann, Michael Baeriswyl, and Martin Jaggi. Simple unsupervised keyphrase extraction using sentence embeddings. 10 2021. doi:10.18653/v1/K18-1022

[3] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength Natural Language Processing in Python, 2020. URL: https://doi.org/10.5281/zenodo.1212303,doi:10.5281/zenodo.1212303.

[4] I. S. Jacobs and C. P. Bean, “Steven Loria. textblob documentation. Release 0.15, 2, 2018.

[5] Pierre Lison and Jörg Tiedemann. Opensubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios.

[6] Radim Rehurek and Petr Sojka. Gensim–python framework for Piperidis, editors, Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016. European Language Resources Association (ELRA), 2016. URL: http://www.lrec-conf.org/ proceedings /lrec2016/ summaries / 947.html.

[7] Layla Oesper, Daniele Merico, Ruth Isserlin, and Gary Bader. vector space modelling. NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic, 3(2), 2011.

[8] Y. Nakatani Shuyo. Language detection library for java.2010. URL: http://code.google.com/p/language-detection/.

[9] Jyotika Singh. jsingh811/pyyoutubeanalysis: Youtube data requests and natural language processing on text, 2021. URL: https://zenodo.org/record/5044556,doi:10.5281/ZENODO.5044556.

[10] Wikipedia contributors. Natural language processing — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Natural_language_processing&oldid = 1030186679, 2021. [Online; accessed 25-June-2021].