

Sentiment Analysis of Social Media Platforms: Concepts, Techniques, Challenges

Prof. Manish K. Hadap¹, Mr. Anurag Zawar², Mr. Himanshu Saurkar³, Mr. Pawan Dawake⁴,
Ms. Pragati Likhkar⁵, Ms. Janhvi Ambade⁶ and Mr. Swapnil Patil⁷

¹⁻⁷Yeshwantrao Chavan College of Engineering, Nagpur, India

Email: m.hadap@gmail.com, {anuragzawar, himanshusaurkar9, psdawake, pragatilikhar87, janhviambade25, swapnilrpatil7812}@gmail.com

Abstract— Sentiment analysis is a vital field focused on understanding the spectrum of emotions, opinions, and attitudes expressed across various social media platforms. By analyzing textual content, researchers aim to categorize sentiments as neutral, positive, or negative. This analytical approach holds immense value in assessing consumer perceptions, brand sentiments, and product preferences. With the omnipresence of social media, individuals worldwide have a platform to voice their thoughts and emotions, making it a rich source for sentiment analysis. Monitoring social media activities enables businesses to gauge consumer sentiment and measure brand loyalty effectively. This study specifically delves into the realm of YouTube comments, utilizing sentiment analysis to quantify the sentiments expressed towards videos. By examining the distribution of positive, neutral, and negative comments, researchers aim to discern patterns that may correlate with the popularity or engagement level of videos.

Index Terms— Sentiment Analysis, Lexicon Based approach, Natural language processing, Multilingual sentiment Analysis task, preprocessing task..

I. INTRODUCTION

In today's interconnected society, individuals form judgments and opinions about various aspects of the world around them, including people, products, places, and events. These attitudes, whether positive or negative, can collectively be termed as sentiments. Sentiment analysis, a burgeoning field, delves into automated techniques for extracting these sentiments from written language. The proliferation of social media platforms has led to an abundance of publicly available, user-generated text on the internet, presenting a vast opportunity to gain real-time insights into people's sentiments. One such prominent social media platform is YouTube, renowned for its video-sharing capabilities. Users flock to YouTube to not only watch videos but also to engage with content through likes, dislikes, and comments. The comment section serves as a forum where users express their opinions and emotions related to the video content, making it a rich source for sentiment analysis.

To enhance sentiment analysis capabilities, our objectives focus on leveraging advanced technologies and methodologies. Firstly, we aim to utilize the YouTube API from Google Cloud, facilitating the efficient retrieval of comments from selected YouTube videos. Additionally, implementing advanced data cleaning techniques such as hyperlink removal, special character elimination, and stemming/lemmatization ensures the accuracy of sentiment analysis results. Integration of the TextBlob library further enhances our capabilities, as it employs a lexicon-based approach to discern the emotional tone of comments.

Expanding our analysis reach across diverse social media platforms is another key objective. By developing functionality to analyse sentiment across platforms like YouTube, Reddit, and Facebook, we aim to broaden the scope of our data sources. For instance, employing Reddit post link input enables us to analyse sentiment on Reddit posts, while incorporating the Facebook Graph API facilitates access token generation for analysing comments on Facebook business pages.

Optimizing user experience and accessibility remains paramount in our endeavours. We prioritize creating a user-friendly interface that allows users to easily select analysis options and input relevant data. Automation of comment retrieval and data cleaning processes enhances system efficiency.

II. LITERATURE REVIEW

- In 2020, Keita Fujihira and Noriko Horibe have suggested This sentiment analysis approach leverages a single sentiment dictionary in the user's native language, avoiding full text translations by translating text word-by-word instead. Initially, the text is broken down into individual words through morphological analysis. Each word is then translated into the corresponding word in the native language. The sentiment of each translated word is evaluated using the native language sentiment dictionary, and the overall sentiment of the text is determined from these individual scores. This method minimizes the risk of losing meaningful sentiment expressions and effectively captures the emotional tone of the text. It is versatile across various languages and minimizes performance inconsistencies that typically arise from linguistic variations. Moreover, the method presents each word's sentiment in multiple synonymous forms in the native language, making it easier for users to identify which specific words influences more sentiment
- In 2016, highlighted the crucial role that lexical databases play as references for discerning positive and negative sentiments, indicating that lexicon-based methods rely heavily on detecting sentiment-laden words listed in these databases. A comparative analysis was conducted using a dataset comprising 1,675 sentences from the political news sector, which included 1,137 positive and 538 negative sentences. After preprocessing the data through tokenization, stop word removal, lemmatization, and part-of-speech (POS) tagging with tools like the Natural Language Toolkit (NLTK) and the Stanford POS tagger, the research implemented sentiment analysis. They used SentiWordNet for the lexicon-based approach and Naive Bayes (NB) and Support Vector Machine (SVM) classifiers in WEKA for the machine learning approaches. Among these, the SVM showed the highest F-measure. The study aims to compare these two sentiment analysis methodologies—lexicon-based and supervised machine learning—using a sandbox version of Apache Hadoop and the Stanford coreNLP library on a sample dataset.
- In 2021, developers Issac Lorentz and Gurjiwan Singh engineered a system using Deep Neural Networks to automatically classify YouTube comments into three sentiment categories: positive, negative, or neutral. Their system demonstrated an accuracy rate between 60% and 84%, when benchmarked against evaluations by humans. The system was trained using advanced techniques, including multinomial naive Bayes and SVMs applied to tweets. Furthermore, they recommended leveraging stochastic gradient descent for training linear classifiers dedicated to tweet analysis, and proposed that emojis within tweets could be useful cues for assigning sentiment labels.
- In 2012, Stefan Stieglitz and Linh Dang- Xuan outlined a two-part framework focusing on data tracking and monitoring, as well as data analysis. The first part discusses various methods for tracking and monitoring politically relevant user-generated content across different social media platforms. The second part addresses different analytical techniques tailored for specific purposes and methodologies. Additionally, the framework categorizes social media into three main types: microblogging, social networking sites (SNS), and weblogs. While there are many platforms that fall into the microblogging or SNS categories, they often differ significantly in their functionalities, target audiences, and objectives. The analysis specifically highlights YouTube and Facebook as leading examples of these categories.
- In 2017, researchers S. L. Lo, E. Cambria, and R. Chiong examined the different multilingual methods used in determining subjectivity and polarity, focusing particularly on languages with limited resources. They found that the prevailing strategies in multilingual research for both subjectivity and polarity analyses include lexicon-based, corpus- based, or translation-based methods, although some studies are shifting towards concept- based and sentic-based approaches. Sentic computing integrates common-sense reasoning to define emotional content linked to entities such as objects, actions, events, and individuals.
- In 2019, Zhao et al have developed a novel image-text consistency driven multimodal sentiment evaluation model initially investigated the relationship between text and images. Subsequently, a multi-modal adaptive sentiment analysis model was developed. Utilizing the established SentiBank framework, mid-level visual

features were extracted and used to represent visual concepts by combining various attributes such as social, textual, and visual features to develop a machine learning model. This proposed model demonstrated superior performance when benchmarked against conventional models.

- In 2019, Park and colleagues crafted a semi-supervised model aimed at addressing challenges posed by documents with incomplete sentiment data. This model not only accounted for the partial data but also preserved the local structures derived from actual datasets. Tested on real-time datasets, the model demonstrated effective performance.
- In 2019, researchers Vashishtha and Susan developed a method to analyse sentiment in social media posts using a novel set of fuzzy rules that incorporated multiple datasets and lexicons. Their approach merged Word Sense Disambiguation and NLP techniques with an innovative unsupervised fuzzy rule-based system to classify comments into negative, neutral, and positive categories. The evaluation involved three sentiment lexicons, four established models, and nine publicly accessible Twitter datasets. The results indicated that this new method achieved superior outcomes.

III. METHODOLOGY

A. Using VADER as a Sentiment Analysis Model

Utilizing Vader as a sentiment analysis model presents a promising avenue for our research project centered on analysing sentiments within social media platforms. Vader, a well-known tool in the field, boasts a pre-trained sentiment lexicon that forms a sturdy foundation for deciphering the emotional undertones embedded in user comments across diverse platforms like Facebook, Reddit, and YouTube. The potential of Vader extends beyond its out-of-the-box capabilities. By engaging in a process known as fine-tuning, we can tailor the Vader model to better understand the intricacies of sentiment expressions prevalent in social media discourse. This involves exposing the model to a corpus of social media data, allowing it to adapt and recognize domain-specific sentiment words and idiomatic expressions commonly used within these online communities.

One of the most significant advantages of leveraging Vader as a transfer learning model lies in its ability to expedite the training process. Unlike starting from scratch, where building a sentiment analysis model requires a substantial amount of annotated social media data, Vader's pre-existing lexicon significantly diminishes this need. Instead, we can allocate our resources towards curating a smaller, yet more targeted dataset for fine-tuning purposes.

This streamlined approach not only accelerates model development but also ensures that our analysis remains attuned to the unique characteristics of social media communication.

Nevertheless, it's imperative to approach this method with a critical eye towards its limitations. While Vader's lexicon serves as a valuable asset, it's essential to recognize that it may harbor biases inherent in its construction. These biases could potentially skew the model's interpretation of sentiments, particularly towards certain linguistic patterns or cultural nuances. Moreover, Vader's classification scheme, which predominantly categorizes sentiments into simplistic labels like positive, negative, or neutral, might overlook the nuanced spectrum of emotions prevalent in social media conversations.

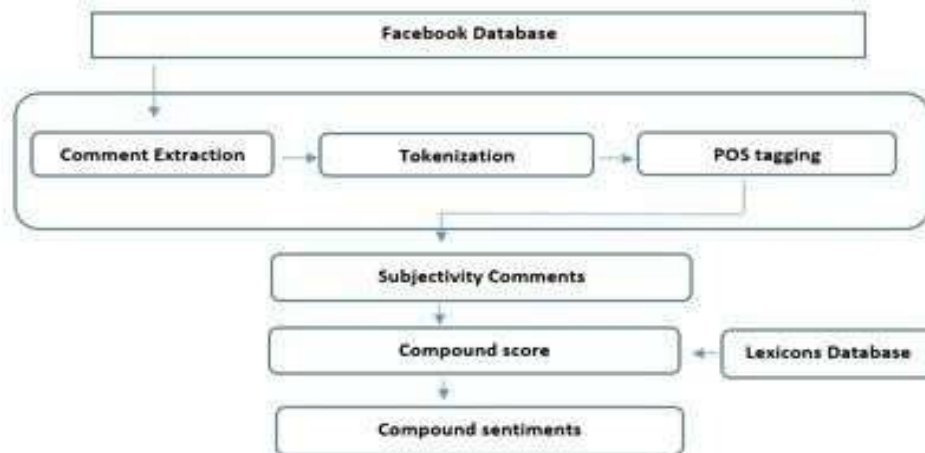


Figure 1. Facebook Database Flow Diagram

Navigating these challenges requires a thoughtful approach. To mitigate biases, present in Vader's lexicon, we must implement strategies that promote inclusivity and fairness in sentiment analysis. This might involve augmenting the lexicon with additional sentiment words or refining the model's algorithms to better account for diverse linguistic variations. Similarly, to address the issue of oversimplification, we can explore methodologies that enable the model to discern subtler emotional nuances, thereby enriching the depth of our sentiment analysis.

The integration of Vader as a transfer learning model into our research project signifies a strategic choice geared towards maximizing efficiency and accuracy in sentiment analysis. By harnessing the strengths of this established tool while acknowledging its limitations, we pave the way for a more nuanced understanding of sentiments expressed within the dynamic landscape of social media. Through continual refinement and adaptation, we can harness the full potential of Vader to glean valuable insights into the ever-evolving realm of online discourse.

B. Preprocessing of Comments

- **Tokenization:** The cornerstone of our social media sentiment analysis project, lays the groundwork for all subsequent text processing tasks. It involves breaking down the raw text data from social media posts, comments, and videos into meaningful units – the tokens. These tokens can be words, numbers, punctuation marks, emojis, or any other element deemed relevant for sentiment analysis.
- **Lemmatization:** Our social media sentiment analysis project can benefit significantly from incorporating lemmatization as a text pre- processing technique. Lemmatization, a crucial step in Natural Language Processing (NLP), aims to group inflected words with their base or dictionary form, known as the lemma. This process helps us analyse the sentiment of the core meaning behind a word, leading to more accurate and consistent results.

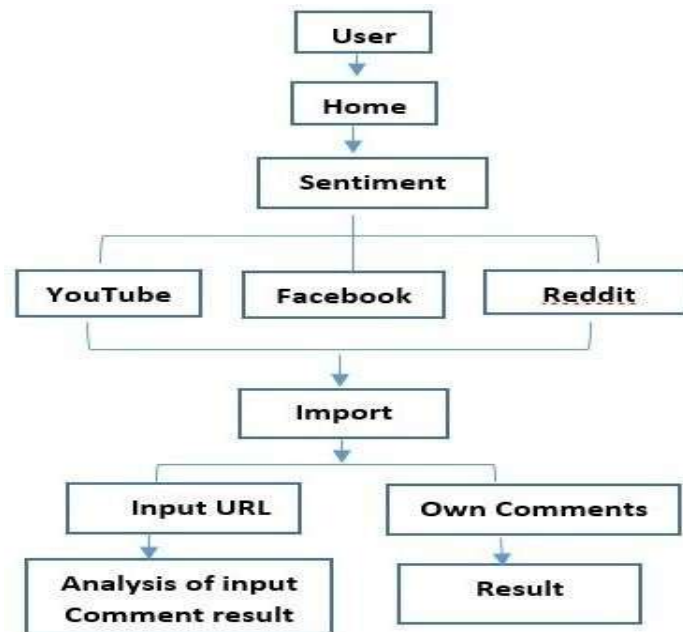


Figure 2. Overall Structure / Flow Diagram

- **Stemming:** Stemming, a text pre-processing technique in our social media sentiment analysis project, aims to reduce inflected words to their base forms (stems). While it offers potential benefits, it's crucial to acknowledge its limitations to ensure effective sentiment analysis.

C. Presenting Sentiments through Interactive Learning

This interactive website offers a user-friendly platform to delve into public sentiment expressed on various social media platforms. By providing options for user-entered comments, live comment scraping, and different social media sources, the website caters to diverse analysis needs. Media sources, the website caters to diverse analysis needs.

D. Question Classification from Social Media Comments using Gradient Boosting Classification

The comments collected from social media and typed by the user are then passed to the question classifier function. This function utilizes a pre-trained Gradient Boosting Classifier (GBC) model specifically designed for question classification. The purpose of this model is to separate questions from the entire list of comments and return the classified questions to the main function for further processing.

Now, let's delve into the detailed explanation of how the model works within this context:

- **Input Data:** The input data consists of comments collected from social media platforms, provided by the user. These comments can include various types of sentences, such as questions, statements, and other text.
- **Preprocessing:** Prior to classification, the comments undergo preprocessing steps to clean and tokenize the text. This step ensures that the input is in a suitable format for the question classifier function.
- **Question Classifier Function:** The question classifier function houses the pre-trained Gradient Boosting Classifier (GBC) model, specifically trained for question classification tasks. Its primary role is to take the preprocessed comments and apply the GBC model to categorize each comment as either a question or non-question.
- **Gradient Boosting Classifier (GBC):** The GBC model used in the question classifier function has been trained on a labeled dataset that includes examples of questions and non-questions. It has learned to recognize patterns and features indicative of questions during its training phase. The model employs TF-IDF vectorization to convert the text of the comments into numerical features, which are then used by the GBC model for classification.
- **Classification Process:** As each comment is passed through the question classifier function, the GBC model predicts whether the comment is a question or not. The classification decision is based on the features extracted from the comment using TF-IDF and the knowledge learned by the GBC model during training.
- **Return to Main Function:** Once all comments have been classified, the question classifier function returns the list of comments along with their corresponding labels (questions or non-questions) to the main function for further analysis or processing.
- **Presenting Imported Comment's Sentiment and Questions through Interactive Website :** We have developed an interactive website using the Django framework, designed to display a list of comments along with their associated sentiment analysis results and extracted questions in a tabular format. Enhancing interactivity, our website includes a Dashboard section where users can visualize the overall sentiment distribution through a pie chart powered by the chart.js module.

The technology stack utilized for website development includes:

Django: A robust Python framework facilitating the creation of dynamic web applications.

HTML: For structuring the content and elements of web pages.

CSS: Employed for styling and enhancing the visual presentation of the website.

JavaScript: Enhancing interactivity and user experience.

Chart.js: Integrated for dynamic and visually appealing data visualization, specifically for the sentiment pie chart.

IV. EXPERIMENTAL RESULTS

A. Scrapping the Comments from Social Media Handles

- **Facebook Comment Scrapping:** Utilized Facebook Graph Api to fetch all the comments of post uploaded in users page. The key required to fetch the comments are post id of which the user is interested to analyze comments, page id, access token provided by graph api to be generated by user.

"Facebook API Limitations: Inability to Fetch Usernames of Comments Due to Privacy Restrictions"

```
url = f'https://graph.facebook.com/v16.0/{page_id}_{post_id}/comments?access_token={access_token}'
response = requests.request("GET", url)
data = json.loads(response.text)
data

{'data': [{'created_time': '2024-03-26T17:08:17+0000',
  'from': {'name': 'Bestlaptops', 'id': '273860125808843'},
  'message': "don't like it. need some modification in it",
  'id': '122105776010250187_914810543771039'}]}
```

Figure 3. Scrapped Comments of Facebook Post

- YouTube Comment Scrapping: Utilized google cloud YouTube API to fetch all the comments by entering the YouTube video URL.

	author	text
0	@skm9593	can we turn it into a json
1	@digitaltechy6217	❤️👍👍 Thanks a ton fir this Video 🙏👍❤️
2	@bertrandlndry	Hello Adam... How can we grab all the comments...
3	@ruari-mcn-fs	Great tutorial thanks mate! Are you a fellow I...
4	@niveditalanksai7686	Is there a way to export the list of comments ...
5	@erramuspe1	Great information, thanks for contribution!!

Figure 4. Scrapped Comments of YouTube Video

- Reddit Comment Scrapping: Used the Praw module to Scrape the comments of provided post URL. The Praw module requires the developer key and secret key provided by reddit to access the comments.

```
my_comm = set()
for comment in post.comments:
    my_comm.add(comment.body)
print(my_comm) #top level comments

["Where you get these service?", "Is it that expensive now? At that cost it is best to group up with some friends and hire a bu
s for a day.\n\nIt used to be 100-150 Rs per day on a bus just before lockdown.", "I only travel by local, best bus, metr
o and taxi. Mumbai is filled with many scams and this city will suck out all your money if you're not careful. Better than Delh
i though.", "I was planning on using these services this week. Can you or anybody else here please elaborate on your experience
while visiting each of the points? Also OP, can you please give a breakdown on what the extra $130 charge is about? Thanks in a
dvance.", "Thanks OP I was planning using this service", "Tourist place like what people all around or smelly sea or shitty str
eets and agnostic narabhai?", "I am coming to mumbai will just use metro local to travel then cab auto for tourist place lets s
ee how much it will cost 🙏👍👍", "I don't know now but I accompanied my relatives in Bombay Darshan around 2018 and we w
ere satisfied with their service. \n\nThey started with a boat ride in Gateway, then Museum > Hanging Garden > Mahalaxi / Haji
```

Figure 5. Comments extracted from Reddit Post

B. VADER Sentiment Result

The project utilizes the Vader sentiment model to analyse comments and generate a dictionary of scores, including negative, positive, and neutral sentiments. The overall sentiment is determined based on the compound score value: scores above 0.5 are positive, below 0 are negative, and between 0 and 0.5 are considered neutral.

```
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer
analyzer = SentimentIntensityAnalyzer()
sentence = "This is best product"
vs = analyzer.polarity_scores(sentence)
print('{:-<80}<'.format(sentence, vs))

This is best product:----- ('neg': 0.0, 'neu': 0.417, 'pos': 0.583, 'compound': 0.6365)
```

Figure 6. Snapshot of VADER Model on Single comment with Score

C. Question Classification Result

```
## TF-IDF based features
vectorizer = TfidfVectorizer(ngram_range=(1,3),
                             min_df=0.001,
                             max_df=0.7,
                             analyzer='word')

X_train = vectorizer.fit_transform(train_text)
X_test = vectorizer.transform(test_text)

y_train = train_df.target.values
y_test = test_df.target.values

gb = GradientBoostingClassifier(n_estimators = 400, random_state=0)
gb.fit(X_train, y_train)
```

Figure 7. Data Preprocessing and Training

	precision	recall	f1-score	support
0	0.94	0.94	0.94	16105
1	0.96	0.96	0.96	26198
accuracy			0.95	42303
macro avg	0.95	0.95	0.95	42303
weighted avg	0.95	0.95	0.95	42303

Figure 8. Accuracy Report of the Model

D. Website



Figure 9. Homepage

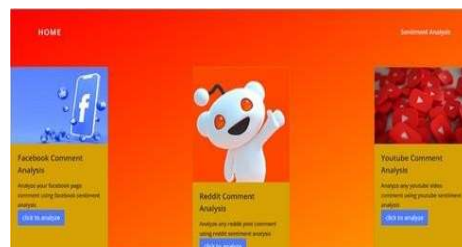


Figure 10. Selection of Social Media for Analysis



Figure 11. Demo of the Outcome



Figure 12. Selection Menu



Figure 13. Input Your Comment



Figure 13. Visualization through Dashboard

V. CONCLUSION

The sentiment analysis project targeting Facebook, Reddit, and YouTube in English and Hindi successfully combines NLP and machine learning to offer insightful sentiment analysis, comprehensive summaries, and intuitive dashboards. Despite challenges such as handling multilingual nuances and social media slang, the project effectively navigates these complexities to deliver a robust analysis tool. This system not only aids in monitoring public sentiment across diverse issues but also paves the way for further advancements, including the integration of more languages, nuanced linguistic features, and real-time data processing.

Ultimately, the project underscores the pivotal role of sentiment analysis in understanding digital era public opinion, showcasing its potential to everything policymaking influence from market trends through technological frameworks. to sophisticated

ACKNOWLEDGMENT

We would like to express our sincere gratitude to everyone who supported and contributed to the completion of this project. Our heartfelt thanks go to Mentor Mr. Manish K. Hadap for his invaluable guidance, insightful feedback and constant encouragement throughout the project.

We deeply appreciate of Yeshwantrao Chavan College of Engineering for providing the resources and support necessary to carry out this project effectively.

REFERENCES

- [1] Multilingual Sentiment Analysis for Web Text Based on Word to Word Translatio n- By Keita Fuijira and Noriko Horibe.
- [2] Sentiment analysis in a cross-media analysis framework- ByYonas Woldemariam.
- [3] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up sentiment classification using machine learning technique," in In ACL Conference on Empirical Methods in Natural Language Processing, 2010, pp. 354-368.
- [4] G.Vinodhini and R.Chandrasekaran, "Sentiment analysis and opinion mining: A survey," International Journal of Advanced Research in Computer Science and Software Engineering., vol. 2, no. 2277 128X, 2012.
- [5] S.Padmaja, S.Sameen, and S.Bandu, "Evaluating sentiment analysis methods and identifying scope of negation in newspaper articles," vol.3 , No.11, no. 11, 2014.
- [6] Olson, D. L, and Delen, Advanced data mining techniques. Dursun:Springer, 2008.
- [7] S. L. Lo, E. Cambria, and R. Chiong, "Multilingual sentiment analysis: from formal to informal and scarce resource languages," Artificial Intelligence Review 48(4), pp. 499-527, 2017.
- [8] K. Dashtipour et al., "Multilingual Sentiment Analysis: State of the Art and Independent Comparison of Techniques," Cognitive Computation, vol. 8, pp. 757-771, 2016.
- [9] Stieglitz, S. & Dang-Xuan, L. Social Media and Political Communication: A Social Media Analytics Framework (Springer, 2012). <https://doi.org/10.1007/s13278-012-0079-3>.