ICEST-2K24 [International Conference on Engineering, Science and Technology]
In Association with
International Journal of Scientific Research in Science, Engineering and Technology
Print ISSN - 2395-1990 | Online ISSN : 2394-4099 (www.ijsrset.com)

Opinion Mining on YouTube Comments

Adarsh Umadi, Abrar Kadri, Hamza Shaikh, Jyotsana Dhekale, Dr.Rupesh Mahajan
Dr. D.Y. Patil Institute of Technology, Pune, Maharashtra, India

ABSTRACT
This paper delves into the realm of opinion mining on YouTube, a platform brimming with user-generated content ripe for sentiment analysis. By harnessing the power of machine learning, we aim to develop a system capable of extracting and classifying opinions expressed within YouTube comments. This will involve constructing classifiers that not only categorize sentiment (positive, negative, or neutral) but also discern the type of comment itself. In order to do sentiment analysis, our approach is to explore how machine learning (ML) and natural language processing (NLP) interact. How to codify human language using specific NLP tools, how to transfer data to meaningful conclusions using those tools, and how MI leverages Python for sentiment analysis. The ultimate goal is to unlock valuable insights from the vast ocean of YouTube commentary, enabling us to gauge public perception on a range of topics and fostering a deeper understanding of audience engagement.

Keywords— Sentiment Analysis, Machine learning, YouTube comments, Natural language processing, emojis

I. INTRODUCTION

A. History
Opinion mining, another name for sentiment analysis, has changed dramatically as social media platforms have grown in popularity. The social media environment has changed significantly, with platforms like Facebook, Instagram, YouTube, WhatsApp, and Twitter emerging as significant hubs for audience engagement and user-generated content. These platforms produce enormous volumes of textual data every day, encapsulating a diverse spectrum of human sentiments and perspectives. Recent years have seen a surge in sentiment analysis research due to developments in natural language processing (NLP) and the accessibility of extensive annotated datasets. But the lack of large-scale, emotion-labeled image datasets has hindered progress in the field of visual sentiment analysis, especially when it comes to emojis and photos. In order to overcome this, cross-domain transfer learning techniques have been used, including pre-training on big datasets like ImageNet and fine-tuning on smaller sentiment datasets.

B. Motivation
The primary motivation behind these studies is to harness the power of sentiment analysis to better understand public perception and individual emotions expressed on social media platforms. Understanding user behavior, emotions, opinions, and experiences can be gained greatly by analyzing sentiments in posts and comments. One way that producers, advertisers, and researchers can measure audience engagement and

response to content is by analyzing the attitudes expressed in YouTube comments. This information is especially helpful for market research and decision-making. The goal goes beyond creating strong algorithms that can recognize general sentiment (good, bad, or neutral) and can also distinguish between the types and degrees of complexity of emotions expressed by text and emojis.

C. Basic Observation

Online discussions, such as those found in YouTube comments, are rich in viewpoints and feelings. These insights can be revealed by sentiment analysis approaches, which go beyond simple positive, negative, or neutral responses. We can convey the subtleties of viewers' emotions, from joy and enthusiasm to irritation and rage, by examining both text and emojis. Deep learning models and other sophisticated technologies are needed for this, along with meticulous data preparation to manage the intricacies of handling multiple languages and emojis.By observing how viewers respond to particular content, they can make informed decisions about future video content, advertising tactics, and even market research. Social media data can be messy, emoji use varies across cultures, and current events can influence how people express themselves.

II. LITERATURE REVIEW

In [1] Riza Velio et al. Sentiment Analysis Using Learning Approaches over Emojis for Turkish Tweets. The study evaluated two techniques (fast Text and Bag of Words) for converting emoticons and emojis into numerical data for sentiment analysis (determining whether they are neutral, positive, or negative).

In [2] Anshika Verma et al. Social Media Sentiment Analysis on Twitter Dataset. Insights on using the random forest method and decision tree extraction than the SVM technique.

In [3] Ziad Al-Halah et al. Smile, Be Happy :) Emoji Embedding for Visual Sentiment Analysis. This paper introduces a novel dataset labeled the Visual Smiley Dataset, which is used to train an emoticon-based image embedding algorithm.

In[4] Xiaomi Sun et al. Fine-grained emoji sentiment analysis based on attributes of Twitter users. Developed a fine-grained analysis method of emojis as same emoji will express different emotions based on tweets.

In [5] Yu Mon Aye and Sint Aung. Contextual Lexicon Based Sentiment Analysis in Myanmar Text Reviews. Proposed a system using Lexicon based approach to classify sentiments of food and restaurants reviews domain in Myanmar.

In [6] Dr. Irish C. Juanatas et al. Sentiment Analysis Platform of Customer Product Reviews. A platform that allows businesses to get insights from consumers about their products.

In [7] Mazen M. Hrazi et al. Sentiment Analysis of Tweets from Airlines in the Gulf Region Using Machine Learning. A sentiment analysis method based on machine learning has been used along with multiple supervised learning algorithms.

In [8] Hanif Bhuyan et al. Retrieving YouTube video by Sentiment Analysis on User Comment. Provides quality, relevance and popular YouTube videos based on users' comments.

III.PROBLEM STATEMENT

In the age of pervasive online communication, the imperative to develop highly effective machine learning models for sentiment analysis has become paramount. These models need to be able to correctly identify the

emotions hidden in user comments and emoticons on review and social networking sites. This urgency arises from the issue of managing the growing amount of unstructured data created in the digital realm, which cuts across industrial borders. Sophisticated sentiment analysis techniques are necessary for both businesses and government organizations to obtain accurate insights into public opinion, developing trends, and consumer sentiments. This project's main goal is to deal with the numerous difficulties involved in managing enormous, disorganized data sources.Bydoing this, it aims to provide decision-makers from a variety of sectors with the information and understanding needed to take well-informed, data-driven decisions. Essentially, the goal of this project is to close the gap that exists between the ever-growing landscape of online communication and the vital requirement for precise sentiment analysis. This will improve the capacity to traverse and utilize the abundance of information present in these digital spaces.

IV. PROPOSED SOLUTION

This research article offers a thorough and multifaceted answer to the complex problems associated with sentiment analysis in the context of social media. This strategy includes a number of essential elements, each of which adds to a comprehensive solution:

1. Data Collection: The first step of the process involves gathering data from different social media platforms, where users contribute a large amount of textual data as well as, occasionally, emojis. The first phase entails gathering this raw data in a methodical manner to make sure it is a representative sample of the target domain.
2. Data Preprocessing: Extensive preprocessing is carried out after the data is collected. Emoji extraction, textnormalization, and handling missing values are some of the responsibilities that fall under this stage. Since social media raw data might be noisy and unstructured, these procedures are essential to guaranteeing data consistency and quality.
3. Feature Engineering: This crucial stage involves extracting pertinent features from the preprocessed data. These capabilities could be text-based as well as, most famously, emoji based. To capture the richness of emotional expression in text, methods such as sentiment scores associated with emojis, word embeddings, and TF-IDF (Term Frequency-Inverse Document Frequency) are used.
4. Machine Learning Models: The research paper emphasizes the utilization of machine learning models for sentiment analysis. Choosing the right models, training them on the prepared dataset, and thoroughly assessing their performance are all steps in the process. Ensuring accurate sentiment classification and grading is the goal.
5. Emojis in Real-World Applications: This research is unique in that it examines emojis' function in sentiment analysis in great detail. Emojis are emotive graphic expressions that are becoming more and more common in digital communication. It is important to comprehend how they affect sentiment, and the research sheds light on this aspect. Furthermore, the study highlights the adaptability and usefulness of the created sentiment analysis tools by proposing real-world uses for sentiment analysis, such as business intelligence and website reviews.
6. Continuous Improvement:The research supportscontinuous improvement because it acknowledges that sentiment analysis and natural language processing arealways changing fields. In order to improve the models and procedures in light of new trends and advancements in the industry, feedback loops are

used. Maintaining therelevance and efficacy of the established tools requireskeeping up with technological improvements in sentiment analysis.

7. Enhanced Decision-Making: Enhancing decision-making and informed action across a wide range of disciplines is the ultimate goal of this approach. The methodology offers businesses, governments, and other stakeholder's significant insights into consumer attitudes, trends, and public opinion through accurate sentiment classification and efficient data processing. This gives them the ability to make data-driven decisions that are in sync with how online communication is evolving constantly.

In simple terms, the approach offered in this research study provides a flexible and methodical framework for sentiment analysis, tackling the particular difficulties brought about by social media and online communication. In an increasingly digital world, it helps to make decisions that are more informed and more successful by bridging the gap between unstructured data and actionable insights.

V. SYSTEM ARCHITECTURE

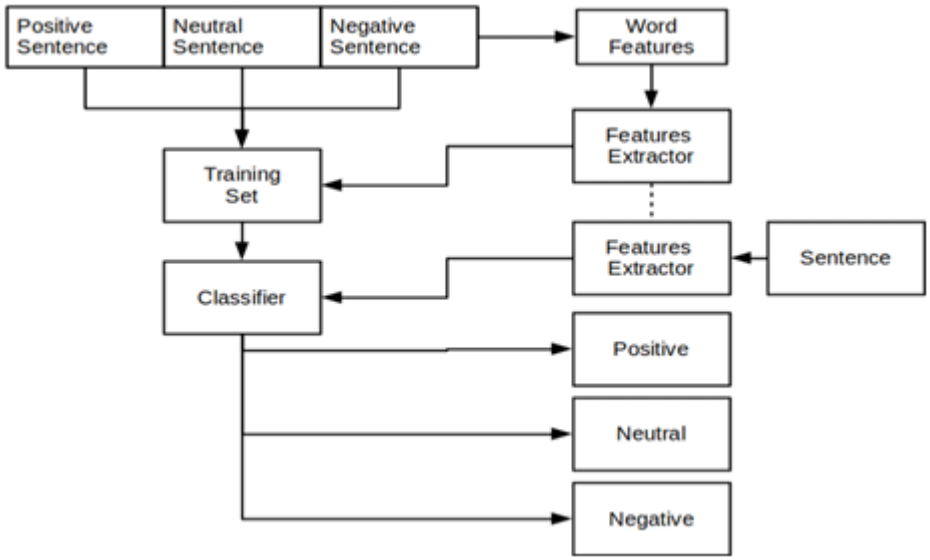


Figure 1.SystemArchitecture

The sentiment analysis system architecture consists of various major components and activities. Initially, input data consisting of positive, neutral, and negative statements are gathered. Each sentence is processed to extract word features using a module called the Features Extractor, which turns raw sentences into useable formats like vectors. These features from the labeled sentences form a training set with both features and their labels. The classifier is then trained on this training set, using machine learning algorithms such as Naive Bayes, Regression or Natural Language Processing to identify patterns associated with each sentiment category. When categorizing a new sentence, the Features Extractor extracts its features, which are subsequently input into the learned classifier. The classifier predicts the emotion of the incoming text using the learnt patterns and returns one of three sentiment labels: positive, neutral, or negative. This architecture effectively manages the entire process, from sentiment classification to text analysis. Deep learning algorithms methodically evaluate each sentence, resulting in an overall sentiment intensity rating. This methodical process assures accurate evaluation of emotional content, making it important for sentiment quantification and nuanced public opinion analysis.

VI. METHODOLGIES

A. Lexicon-based Approach:

Lexicon-based sentiment analysis is a rule-based approach that relies on pre-built dictionaries of words with sentiment scores. These dictionaries, also called lexicons or sentiment lexicons, map words to their emotional polarity (positive, negative, or neutral) and sometimes include an intensity score.

A lexicon, also called a sentiment dictionary or opinion lexicon, is essentially a list of words with assigned sentiment values. These values can be:

- Positive (e.g., happy, love)
- Negative (e.g., sad, hate)
- Neutral (e.g., the, and)

Sometimes, lexicons might include additional information like intensity levels (weak vs strong positive/negative). Sentiment scores are allocated to every word in the lexicon.This score can contain intensity levels (e.g., "great" = 2, "bad" = -1.5) or be binary (positive = 1, negative = -1, neutral = 0).

Sentence Sentiment Score = Σ (Word i sentiment score)

- Σ (sigma) represents the summation over all words (i) in the sentence.

B. Naive Bayes:

A well-liked supervised learning technique for classification problems, Naive Bayes works especially well for text classification. It uses the Bayes theorem to estimate probabilities and create forecasts. Feature independence is the fundamental tenet of Naive Bayes, hence the name. It makes the assumption that any feature that affects the categorization is unrelated to every other feature. Actually, features may be connected (for example, the terms "hot" and "sunny" frequently occur together). But this simplicity makes for an algorithm that is quick and effective.

This probabilistic classifier calculates the probability (P) of a text document (d) belonging to a sentiment class (c) based on the presence of words (w):

P(c|d) = P(d|c) * P(c) / P(d)

- P(c|d): Probability of class (sentiment) c given document d (posterior probability)
- P(d|c): Probability of document d given class c (likelihood)
- P(c): Prior probability of class c (independent of the document)
- P(d): Total probability of the document (usually constant)

Each sentiment class's likelihood is determined via Naive Bayes, which then allocates the document to the class with the highest probability.

C. Recurrent Neural Networks (RNNs):

When it comes to sentiment analysis, recurrent neural networks (RNNs) are an extremely useful tool, especially when working with text sequences such as phrases or reviews. Sequential data is a strong suit for RNNs, as opposed to autonomous feature analysis found in algorithms such as Naive Bayes. Word by word, they read the text, taking into account the connections and sequencing of the words. This is important for sentiment analysis since the sentiment of a word can be affected by the words around it. Below is the process-flow of RNN:

- Information is processed through hidden layers in an RNN's structure, which resembles a loop.

- The RNN receives a word as input at each step, coupled with the hidden state from the previous step that represents the context that has been seen thus far.
- After processing this data, the RNN modifies the hidden state, essentially retaining the knowledge it acquired from earlier words.
- As the sentence progresses, the RNN can develop a comprehension of the sentiment by using this updated hidden state to process the subsequent word.

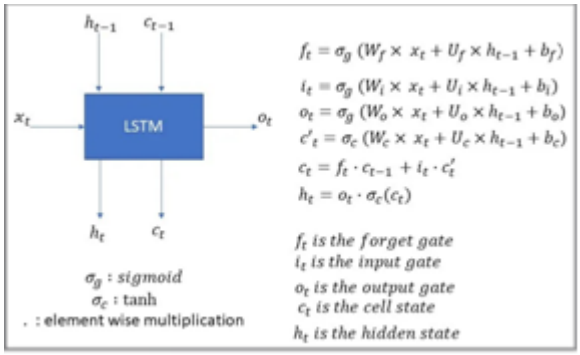
A basic RNN unit involves an activation function applied to a combination of the current input (word embedding) and theprevious hidden state (capturing context):

$$h_t = f(W_h * h_{(t-1)} + W_x * x_t + b)$$

- h_t: Hidden state at time step t (output of the RNN unit)
- f: Activation function (e.g., tanh, Re-LU)
- W_h: Weight matrix for hidden layer connections
- W_x: Weight matrix for input layer connections
- x_t: Input vector at time step t (word embedding)
- b: bias vector

D. Long Short-Term Memory (LSTM) networks:

Recurrent neural networks (RNNs) with Long Short-Term Memory (LSTM) networks are very good at sentiment analysis, especially when working with lengthy text sequences. Long-term dependencies in sequences are difficult for typical RNNs to learn because to the vanishing gradient problem, which LSTMs solve. Because of this, LSTMs are especially useful for sentiment analysis,where it's critical to comprehend the context of a statement.



Process-flowof LSTM in Sentiment Analysis:

- Data preprocessing: Tokenization (word-by-wordbreakdown) and cleaning of text data are used to provide numerical representations that are appropriate for the model.
- LSTM Model Construction: First, an embedding layer is used to convert words into vectors. Next, LSTM layers areapplied to process the sequence. Finally, final layers are applied to predict sentiment, which is often positive, negative, or neutral.
- Training the Model: The LSTM model learns thecorrelation between text sequences and sentiment labels by means of labeled sentiment data.
- Sentiment Prediction: Following training, the model is able to forecast the sentiment of fresh, unobserved textual input.

E. Sentiment Classification:

The final output of the LSTM layer (hidden state sequence) is fed into a dense layer with a soft-max activation function:

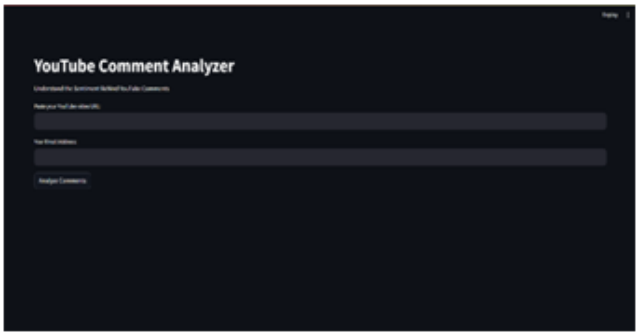
$$P(c \mid h_t) = \text{soft-max}(W_s * h_t + b_s)$$

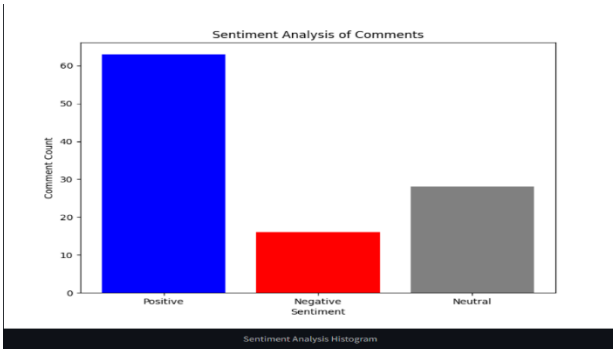
- $P(c \mid h_t)$: Probability of sentiment class c given the final hidden state (h_t)
- W_s: Weight matrix for the sentiment classification layer
- b_s: Bias vector for the sentiment classification layer

The most likely sentiment for the input text can be predicted by the modelthanks to the soft-max function, which provides aprobability distribution over sentiment classes (such as positive, negative, and neutral).

VII. EXPECTED RESULTS

By examining text and emojis, a YouTube comment analyzer can offer comprehensive insights. Positive, negative, or neutral sentiment may be determined generally from the comments, and it can also evaluate sentiment within individual comments and correlate the use of emojis with sentiment. It also has the ability to recognize important conversation starters and monitor the progression of talks over time. Emojis have an impact on comment discussions, and engagement research can highlight viewers who are excited. You can develop upcoming content and discover what your audience responds to by identifying both good and negative feedback. To further improve your comprehension of viewer response, the analyst can even create a personalized emoji sentiment lexicon over time. Your YouTube channel strategy can be guided by this thorough analysis, which can also help with content production and audience interaction.





VIII. FUTURE SCOPE

Future work will concentrate on improving scalability for analyzing large datasets, addressing data quality issues with sophisticated cleaning approaches, and expanding the tool's utility by adding multilingual analytic capabilities. Explainable AI integration will also provide users a better understanding of the logic underpinning sentiment and subject categorization.We can fully realize the potential of audience insights concealed in online comments by consistently improving and refining this YouTube comment analyzer. This will promote a more data-driven approach to social research, customer engagement, and content production.

IX. CONCLUSION

This project looked at creating a YouTube comment analyzer, a tool that extracts insightful information from user comments by using natural language processing techniques. In order to categorize comments as favorable, negative, or neutral, the analyzer used sentiment analysis. This allowed it to display the general distribution of sentiment and identify the main emotional drivers of the conversation.

Additionally, prevalent themes mentioned in the comments were discovered by topic modelling approaches, giving researchers, businesses, and content creators a better grasp of the talking points and audience attention. For a variety of stakeholders, the created tool is an invaluable resource. Businesses may obtain useful consumer sentiment data, academics can examine public opinion and online conversation dynamics, and content makers can use the insights to fine-tune their plans.Even though the project's primary goals were met, there is always space for development.

X. REFERENCES

[1]. Adarsh Umadi, Abrar Kadri, Jyotsna Dekhale, Hamza Shaikh, "Sentiment Analysis using Machine Learning Algorithms", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395- 1990, Volume 10 Issue 6, pp. 64-70, NovemberDecember 2023.Journal URL : https://ijsrset.com/IJSRSET2310567

[2]. Pak A, Paroubek P. Twitter as a corpus for sentiment analysis and opinion mining. In Proc. LREC, May 2010.

[3]. Pontiki, Maria, et al. "SemEval-2016 task 5: Aspect based sentiment analysis." ProWorkshop on Semantic Evaluation (SemEval-2016). Association for Computational Linguistics, 2016.

[4]. Fang, Xing, and Justin Zhan. "Sentiment analysis using product review data." Journal of Big Data 2.1 (2015).

[5]. Shikha Tiwari, Anshika Verma, Peeyush Garg, Deepika Bansal. Social media sentiment analysis (2016).

[6]. Alameda-Pineda, E. Ricci, Y. Yan, and N. Sebe. Recognizing emotions from abstract paintings using non-linear matrix completion. In CVPR, 2016.

[7]. S. Cappallo, S. Svetlichnaya, P. Garrigues, T. Mensink, and C. G. Snoek. New modality: Emoji challenges in prediction, anticipation, and retrieval. IEEE Transactions on Multime- dia, 21(2):402–415, 2018.

[8]. Ziad al-halah, A. Aitken, W. Shi, J. Caballero. Smile, Be Happy :) Emoji Embedding for Visual Sentiment Analysis. 019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW).

[9]. K.-C. Peng, T. Chen, A. Sadovnik, and A. C. Gallagher. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In CVPR, 2015.

[10]. S.Santhanam, V.Srinivasan, S.Glass,andS.Shaikh. I stand with you: Using emojis to study solidarity in crisis events. In Proceedings of the 1st International Workshop on Emoji Understanding and Applications in social media, 2018.

[11]. Theodora Koulouri, ROBERT D.MACREDIE,and DAVID OLAKITAN, "Chatbots to Support Young Adults' Mental Health: An Exploratory Study of Acceptability," Department of Computer Science, Brunel University, July 2022.

[12]. M. Z. Asghar, A. Khan, S. Ahmad, M. Qasim, and I. A. Khan, "Lexicon-enhanced sentiment analysis framework using rule-based classification scheme," PLOS ONE, vol. 12, no. 2, p. e0171649, Feb 2017, doi: 10.1371/journal.pone.0171649.

[13]. B. Liu, "Sentiment analysis and subjectivity," in Handbook of Natural Language Processing, Second Edition., 2010, pp. 627–666.

[14]. J. Bhaskar, Sruthi K, and P. Nedungadi, "Enhanced sentiment analysis of informal textual communication in social media by considering objective words and intensifiers," in International Conference on Recent Advances and Innovations in Engineering (ICRAIE-2014), Jaipur, India, May 2014, pp. 1–6.

[15]. S. M. Mohammad, F. Bravo-Marquez, M Salameh, and S. Kiritchenko, "Semeval-2018 task 1: Affect in tweets," Proceedings of the International Workshop on Semantic Evaluation (SemEval-2018), New Orleans, LA, USA, June 2018.

[16]. Barman, Utsab& Das, Amitava & Wagner, Joachim & Foster, Jennifer. (2014). Code Mixing: A Challenge for Language Identification in the Language of Social Media. 10.13140/2.1.3385.6967.

[17]. B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up sentiment classification using machine learning techniques," in In ACL Conference on Empirical Methods in Natural Language Processing, 2010, pp. 354-368.

[18]. C. Musto, G. Semeraro, and M. Polignano, "A comparison of lexicon-based approaches for sentiment analysis of microblog posts," on 8th International Workshop on Information Filtering and Retrieval, 2014.

[19]. 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.

[20]. Do, P. Prasad, A. Maag, and A. Alsadoon, "Deep Learning for Aspect-Based Sentiment Analysis: A Comparative Review," Expert Systems with Applications, vol. 118, pp. 272-299, March 2019.