iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** I **Month of publication:** January 2025

DOI: https://doi.org/10.22214/ijraset.2025.66475

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue I Jan 2025- Available at www.ijraset.com

YouTube Comments Sentiments Analysis

Usha Krishna, Sandesh Srivastav, Sandhya Gupta, Sarvesh Chandra Mishra, Vivek Verma
JSSATEN, Noida, Uttar Pradesh, India

Abstract: ***It's the YouTube Comments Sentiment Analysis Web Application, an advanced content tool for content creators that's designed to analyze viewer response on social media, mostly on YouTube. This tool helps in the extraction process and subsequent analysis and classification of comments to give better insight into audience sentiment. The application utilizes the YouTube Data API to download particular video comments on their URLs. The application uses advanced techniques of Natural Language Processing to classify comments into three sentiment classes: positive, negative, and neutral. Results are detailed files organized into Excel and mailed to users. Additionally, an interactive HTML table shows the distribution of sentiment for a quick overview. This friendly tool allows content creators to track audience engagement, see trends in viewer feedback, and make data-driven decisions on improving content quality and relevance.***

The application saves time by automating the comment anal- ysis process. It delivers actionable insights, allowing creators to build stronger connections with their audience and to better optimize their content strategies to enhance engagement and satisfaction.

Keywords: YouTube Comments Analysis, Sentiment Classification, NLP, Viewer Feedback, YouTube Data API, Automated Extraction, Sentiment Visualization, Audience Engagement, Content Optimization.

I. INTRODUCTION

YouTube has become an unbounded treasure of user storage and has been attracting millions of users and creators alike. Analyzing the sentiment narrated in comments, titles, and de- scriptions will provide deep insights regarding user preferences and engagement levels. YouTube sentiment analysis project revolves around advanced natural language processing NLP methods that would classify the sentiments under categories such as positive, negative, and neutral maintained with a focus on analyzing emotions such as ”anger”, ”joy”, or ”sadness”. The project addresses some of the challenges that come with unstructured and informal text data and seeks to improve content analysis and enhance the user experience.

The most popular form is Real-world challenge-solving, which has emerged recently as the importance of senti- ment analysis grows. For example, Nirmalya Thakur et al. [2024] have presented a very robust dataset of 4011 videos on YouTube and TikTok, including emotional labels for sentiment analysis with tools like VADER, TextBlob, and DistilRoBERTa-base. Fine- grain emotion classification re- veals quite a large number of video titles and descriptions bearing a neutral sense.

Other names in this category include Ceren Cubukcu Cerasi and Yavuz Selim Balcioglu [2023]; and Singh and Tiwari [2021]. Their dedicated research works reflect the reality that machine learning techniques like LSTM and SVM would do wonders in the analysis of YouTube comments and other informal writing, negation, and linguistic ambiguity. Such things have instruments craved to implement complex models and preprocessing in order to attain high accuracy in sentiment analysis, especially with multilingual datasets.

Digital marketers are not onlyifications for content-wise recommendations and categorization systems but also evident improvements as discussed in Aditya Baravkar et al. [2020].

II. LITERATURE REVIEW

Nirmalya Thakur et al. [2024] developed a dataset of video clips labeled under sentiment analysis regarding the 2024 measles outbreak, numbering a total of 4011 videos obtained from sites such as YouTube, TikTok, and many more. The records include URLs, titles, descriptions, and publication dates, out of which 48.6%, were drawn from YouTube, and 15.2% are TikTok videos. Using VADER for sentiment anal- ysis, TextBlob for subjectivity analysis, and DistilRoBERTa- base for fine-grain sentiment analysis emotions classified into positive, negative, or neutral were also derived as anger, joy, or sadness.

A greater percentage of titles (62.78%) and descriptions (40.46%) were neutral. This dataset closes the gaps in research because it merges sentiment and emotion classification in the sense of the FAIR principles of accessibility and usability for video-based social media analytics.

Ceren Cubukcu Cerasi and Yavuz Selim Balcioglu [2023]performed the sentiment analysis on comments col- lected from YouTube videos where ChatGPT is mentioned.

They also classified these comments based on their polari- ties. For the analysis, they selected 1000 comments randomly from the top 100 YouTube videos. They also used lexicon approaches like WordNet to determine the polarity of emotion.

The study employed Long Short-Term Memory (LSTM) to classify the comments with high precision and recall scores appropriate to specific categories of videos such as News and Entertainment.

The last thing that this study reveals is the complexities in the use of informal writing and negation analysis which stresses the necessity for much more sophisticated social lexicons and event classification. This research will open the museum gates to insights into user perceptions concerning Chat GPT and put forth recommendations for betterization of future lexicon validation for sentiment analysis studies.

Rawan F. Alhujaili and Wael M.S. Yafooz [2021] reviewed sentiment analysis (SA) techniques used for YouTube com- ments and classified the types of sentiment into three levels: simple, complex, and advanced, all based on machine learning (ML) and deep learning. They mentioned how preprocessing steps like tokenization and normalization are important for accuracy.

The principal models discussed include Na¨ıve Bayes (NB), Support Vector Machines (SVMs), and Convolutional Neural Networks (CNNs). For instance, Krouska et al. [2016] im- proved accuracy using feature selection whereas Al-Tamimi et al. [2017] accomplished an 88.8% score of F-measure using SVM-RBF for Arabic comments.

Bhuiyan et al. [2017] performed classification using Sen- tiStrength and achieved an accuracy of 75.4%. The study stresses the need for further research over the non-English datasets, especially Arabic, to improve video retrieval as well as to use it to maintain user activity.

Singh and Tiwari [2021] have used six machine learning algorithms- Nayve Bayes, SVM, Logistic Regression, Deci- sion Tree, KNN, and Random forest. They have done in- depth content analysis for the work that has been carried out on YouTube comments.

About 1500 comments have been annotated, which are separated into distinct categories such as positive, negative, or neutral, and applied preprocessing like lemmatization, tok- enization, and removing stop words.

Among the classifiers, SVM proved to be the most accurate, while n-grams combined with feature selection improved performance for DT and RF.

This research demonstrated the influence of real-world events on effect and also highlighted how effective ML is with respect to analyzing YouTube comments.

Aditya Baravkar et al. [2020] have devised a system of sentiment analysis for better discovery of educational YouTube videos. The model analyzes the sentiments of comments, counts of likes, views, and top comment sentiments with the classifying Regression method.

The web application ranks videos through a customized sort algorithm, thereby delivering high-quality content and lessening the search time of users. The framework exhibits an opportunity to expand beyond categories and serves as a solid recommendation model for YouTube.

Mohd Majid Akhtar [2019] has succeeded in developing a sentiment analysis model capable of classifying YouTube comments into positive, negative, or neutral sounds, based on the functionality of TextBlob, by giving polarity scores that range from -1 to +1.

The methodology includes extracting the data through the video IDs, converting it to CSV files, and analyzing it by means of sentiment.

The football-themed comments were so processed, which ended up giving the online system a 70% accuracy, 100% precision, and 75% recall performance.

It highlighted the need to improve the classification tech- niques in noisy datasets in addition to some informal lan- guages, context relevance, and noise in datasets needed for more accuracy.

Zulfadzli Drus et al. [2019] conducted a survey on sentiment analysis in social media from the year 2014 until 2019, including lexicon-based and machine learning techniques, with data sourced primarily from Twitter.

Most applications of sentiment analysis have been recorded in business, politics, health care, and disaster response, demon- strating its ability to inform decision-making.

The study emphasized hybridization to estimate better ac- curacy. Further, it called for research in different platforms to create universal models.

TABLE I: Gap Analysis

S.N	Author	Year	Proposed System	Gap
1	Thakur, et al.	2024	This dataset holds a collection of 4,011 videos, comprising coverage on measles outburst for the year 2024 as col- lected from 264 sources which in- clude YouTube, TikTok, Instagram, and Facebook. The dataset contains sen- timents (VADER), fine-grained senti- ments (best-for-its-size DistilRoBERTa- base), and subjectivity (TextBlob) analy- sis.	• The dataset does not contain any videos related to measles outbreaks that hap- pened in the year 2024. • Most datasets lack sentiment and fine- grain sentiment attributes. • Absence of datasets that combine differ- ent sources like social media and news websites.
2	Çubukc¸u Ç erasi, et al.	2023	Sentiment analysis of 1,000 randomly picked YouTube comments on videos of ChatGPT by lexicon-based sentiment analysis of the comments and classifica- tion using long short-term memory.	• Classifying informal language and nega- tion in comments is a real challenge. • Limiting existing emotion lexicons.
3	Rawan Fahad Alhujaili and Wael M.S. Yafooz	2022	Modeller and Lexicon-Based Techniques of Grades Simple, Complex, and Ad- vanced Reviewed. Sentiment Analysis Techniques for YouTube Comments Fo- cusing on Machine Learning and Lexicon- Based Methods, Bringing out the Key Role of Pre-processing in Improving the Classification Accuracy.	• Restricted interest in multilingual sen- timent analysis, especially for less re- sourceful ones. • Inadequate assessment of deep learning models and their comparative advan- tages.
4	Ritika Singh and Ayushka Tiwari	2021	Proposed six machine learning algorithms like NB, SVM, LR, DT, KNN, and RF in sentiment analysis of YouTube comments. The various preprocessing steps applied are stop word removal, lemmatization, and different model evaluation metrics such as F-score and accuracy.	• Very small enough dataset at a size of only 1500 annotated citation sentences. • Doesn't have deep learning models, which, in fact, could surpass traditional classifiers.
5	Aditya Baravkar et al.	2020	Designed a sentiment analyzer app for educational YouTube videos using com- mons, likes, views, and a logistic regres- sion analyzer. It also includes a web ap- plication that the videos on personalized sorting algorithms.	• Focused more on a very small dataset restricted to specific content types (edu- cational videos). • Did not consider the contextual senti- ment for wrong detection of unrelated positive or negative comments.
6	Mohd Majid Akhtar	2019	Developed a sentiment analysis model to classify YouTube comments as positive, negative, and neutral based on TextBlob. Videos on football were chosen, alongside a minimal data sample.	• Limited scope as only 50 comments available. • It follows a rule-based approach using Text blob which has no context for appli- cation and misclassification is common.
7	Zulfadzli Drus et al.	2018	A systematic review of sentiment analy- sis methodologies has been carried out, primarily including methods that use lex- icons and machine learning techniques.	• Predominantly concerned with the En- glish language sentiment analysis and not paying attention to multilingual or low- resource languages.

III. PROPOSED WORK

This easily manageable intelligent system is now going to improve the extraction, analysis and interpretation of audience sentiment through comments on YouTube videos. It applies state-of-the-art machine learning (ML) and natural language processing (NLP) techniques to classify comments into three classes, i.e. positive, negative and neutral. In the end, this system will positively affect content creators, researchers, and marketers; through it, they get the real knowledge to improve ways and strategies to engage with the audience better while understanding their view of the content.

A. Key to the Proposed System

1) Automated Data Retrieval: Insider-like automated retrieval of comments and associ- ated metadata like usernames, timestamps from the YouTube via YouTube Data API or automated webscraping.

2) Comprehensive Pre-Processing of Data: • Involve ad- vanced preprocessing, including noise removal, tokenization, sentiment lexicon application to qualify the input data for analysis.

3) Very Accurate Sentiment Classification: • State-of-the-art sentiment analysis tools like VADER or custom trained ML models to provide sentiment scoring-based tagging-positive, negative, and neutral-for any comment.

4) Beautiful Visual Reporting:

- These are presented in an interactive dashboard with sentiment trends in the forms of bar graphs, pie charts, and other visuals.
- It will also generate downloadable reports such as CSV files for further detailed analysis of the sentiment breakdown.

5) Friendly User Interface:

- Build a web application that responds to both Flask or any similar frameworks that accept video URLs, allowing for the viewing of results without hiccups.
- Send an email with categorized reports directly to the user's inbox.

B. System Objectives

To justify audience analysis: Content creators and re- searchers should understand their audiences better in terms of feedback.

- Content strategy improvement: Give actionable insights for better quality video content, engagement, and reach.
- Accessibility of Data: The system should be intuitive, scalable, and accessible to diverse user groups.

Expected Results

1) Target Audience Engagement: By monitoring both positive and negative sentiments, the content creation will change according to what the target audience expects.

2) Actionable Knowledge in Growth: The deep analysis of sentiment will give data-driven insight to the action that would alter the content and result in loyalty for the audience.

3) Content Discovery: The outcome of sentiment analysis will provide an oppor- tunity for low-rated videos with a high rating and support. The system follows a sequential four-step process:

C. Data Collection (YouTube Comments Scraping)

Choose an appropriate method or tool for scraping YouTube comments, such as using the YouTube API or web scraping libraries like BeautifulSoup or Scrapy or Selenium library of python programming language. Retrieve comments from selected videos, ensuring that the data collected includes relevant metadata (e.g. username and the comments).

1) Preprocessing Data: Clean the scraped data to remove noise, such as HTML tags, emojis, and irrelevant charac- ters.Tokenize the comments into individual words or phrases for further analysis.

2) Sentiment Analysis: Apply sentiment analysis techniques to the sentiment expressed each comment.Choose an appro- priate approach, such as lexicon-based in methods (e.g., using sentiment dictionaries) or machine learning algorithms.Assign sentiment scores (e.g., positive, negative, neutral) to each comment based on the analysis.

3) Deployment: Once the sentiment analysis pipeline is devel- oped, it can be deployed as a service or integrated into existing applications. This module ensures that the sentiment analysis functionality is accessible to end-users through an API, web interface, or command- line interface.

4) Data Flow: In a YouTube comment scraping and sentiment analysis project, the process begins with the user providing input, such as a video URL or a search term, via the user interface. The system then leverages the YouTube API or a web scraping tool to extract comment data from the specified videos.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue I Jan 2025- Available at www.ijraset.com

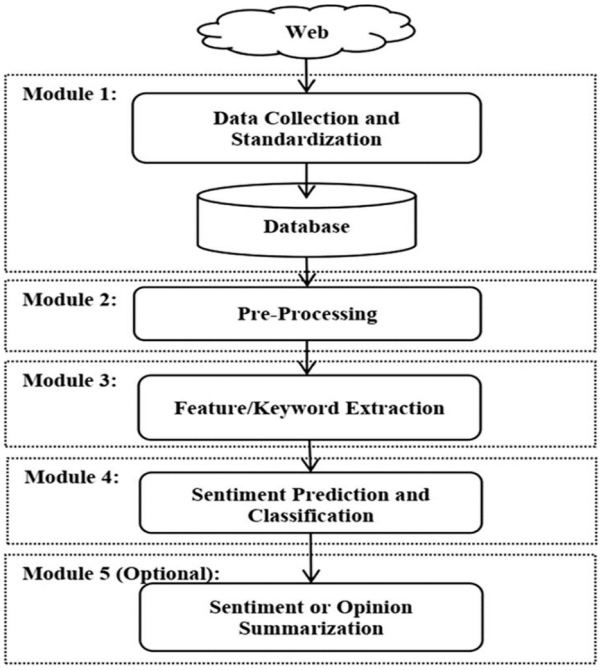


Fig. 1. System Architecture

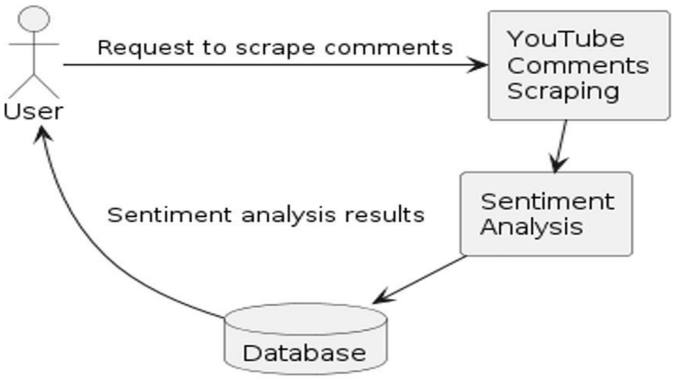


Fig.2. Data Flow

D. Libraries and Algorithms Used

This project uses different machine learning algorithms and develops the functionality of the system with the help of a combination of various Python libraries and tools.The major ones are Flask, pandas, NLTK, Selenium, and numpy each for a specific use within the overall workflow of the YouTube comment scraping and sentiment analysis application.

1) Web Interface Development with Flask: It is the de- velopment of the UI part of the application, being the web application framework pivotal for lightweight development. Even through its minimalist design, it would be easy to implement a web interface where one could type in a URL of a YouTube video to interact with the system by commenting scraping and doing some sentiment analysis on the same.

2) Data Handling and Manipulation using pandas: One of the integral libraries used for the system in handling and processing data is pandas. In particular, pandas is heavily utilized with CSV files, where YouTube comment extraction and sentiment analysis results are stored. pandas simplifies data manipulation tasks, hence making categorization and other analytical commenting easier.

3) Automated Google Comment Extraction Using Sele- nium: Selenium is the most powerful web automation tool that scrapes comments from YouTube video streams. When given a URL of the video, Selenium automatically navigates to the relevant video page and pauses the video from playing to provide scope for simulating scrolling actions for loading the comments section. After that, the usernames and comments can be downloaded into CSV files for further analysis.

4) Sentiment analysis using NLTK and VADER: The system uses the NLTK (Natural Language Toolkit) for all its natural language processing tasks. Most specifically, VADER (Valence Aware Dictionary and Sentiment Reasoner) from NLTK per- forms sentiment analysis on all the extracted comments at present.Every comments is given sentiment scores indicating its positivity, negativity and neutrality. Depending on those scores, classifications into being either positive or negative will be made.

5) Role of numpy: The numpy library for numerical comput- ing is imported in the system code, but there is no explicit reference in the present way for its use with this implementa- tion. Importing points towards future possible use in handling mathematical operations or arrays.

IV. CONCLUSION

This web-based animation tool for visualizing sorting algorithms proved quite effective and beneficial mainly because of the huge inputs put into development. The feedback received from students who used this tool turns out to be mostly in agreement with the earlier studies, which has shown that there is no very great difference in content comprehension through traditional methods against animated tools. It underlines the importance of developing animated presentations and putting them to use in the classroom for enhancing education. As long as the age of JavaScript lasts, this tool will be found relevant and adaptable without the major need of redoing it in the next programming language.

REFERENCES

[1] T. Bingmann. “The Sound of Sorting - ‘Audibilization’ and Visualization of Sorting Algorithms.” Panthemanet Weblog. Impressum, 22 May 2013. Web. 29 Mar. 2017.

[2] http://panthema.net/2013/sound-of-sorting

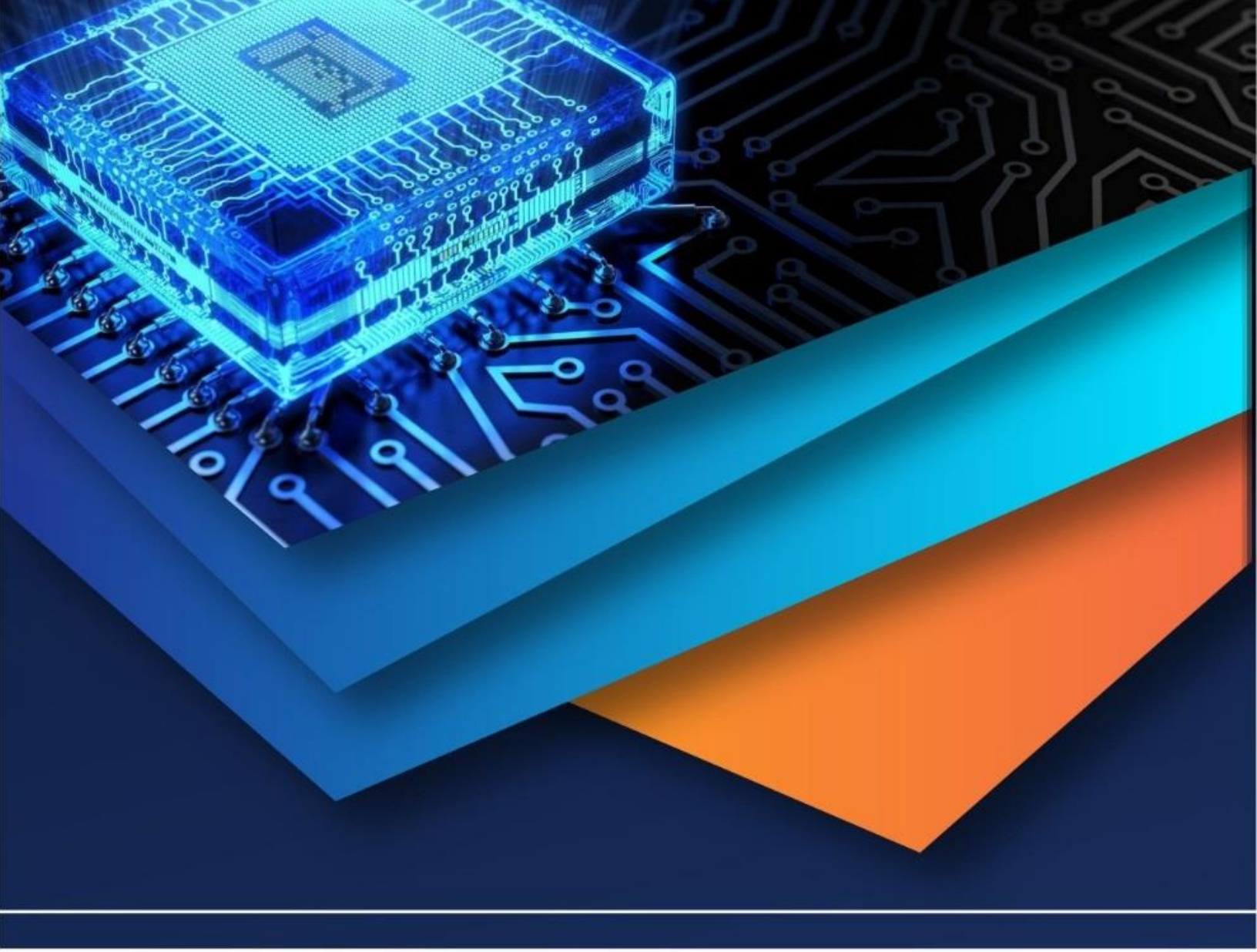
[3] Bubble-sort with Hungarian (“Cs´ang´o”) Folk Dance. Dir. K´atai Zolt´an and T´oth L´aszl´o. YouTube. Sapi- entia University, 29 Mar. 2011. Web. 29 Mar.2017.

[4] Kerren and J. T. Stasko. (2002) Chapter 1 Algorithm Animation. In: Diehl S.(eds) Software Visualization. Lecture Notes in Computer Science, vol 2269. Springer, Berlin, Heidelberg.

[5] Moreno, E. Sutinen, R. Bednarik, and N. Myller. Conflic- tive animations as engaging learning tools. Proceedings of the Koli Calling ’07 Proceedings of the Seventh Baltic Sea Conference on Computing Education Research - Volume 88, Koli ‘07 (Koli National Park, Finland), pages 203-206.

[6] J. Stasko. Using Student-built Algorithm Animations As Learning Aids. Proceedings of the Twenty eighth SIGCSE Technical Symposium on Computer Science Education. SIGCSE ‘97 (San Jose, California), pages 25-29. http://doi.acm.org/10.1145/268084.268091

[7] J. Stasko, A. Badre, and C. Lewis. Do Algorithm Ani- mations Assist Learning?: An Empirical Study and Anal- ysis. Proceedings of the INTERACT ‘93 and CHI ‘93 Conference on Human Factors in Computing Systems, CHI-93 (Amsterdam, 6.. the Netherlands), pages 61-6. http://doi.acm.org/10.1145/169059.169078



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY