# Sentiment Analysis on Youtube Comments

***A project report submitted to MALLA REDDY UNIVERSITY***

***in partial fulfilment of the requirements for the award of degree of***

## BACHELOR OF TECHNOLGY

**in**

## COMPUTER SCIENCE & ENGINEERING (AI & ML)

**Submitted by**

**J. Phani Kumar : 2011CS020158**

**K. Jathin Guptha : 2011CS020166**

**M. Srinath : 2011CS020232**

**M. Vineela : 2011CS020238**

***Under the Guidance of***

**Dr. Sujit Das**

**Assistant Professor**

**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING (AI & ML)**

2023



**COLLEGE CERTIFICATE**

This is to certify that this is the bonafide record of the application development entitled, “**SENTIMENT ANALYSIS ON YOUTUBE COMMENTS**” Submitted byY.PhaniKumar(2011CS020158),K.JathinGuptha(2011CS020166),M.Srinath(2011CS020232),M.Vineela(2011CS020238) B. Tech IV year I semester, Department of CSE (AI&ML) during the year 2023-24. The results embodied in the report have not been submitted to any other university or institute for the award of any degree or diploma.

**INTERNAL GUIDE HEAD OF THE DEPARTMENT**

**Dr .Sujit Das Dr. Thayyaba Khatoon**

**Assistant Professor CSE(AI&ML)**

**EXTERNAL EXAMINER**

# ACKNOWLEDGEMENT

The satisfaction that accompanies the successful completion of any task would be incomplete without the mention of people whose ceaseless cooperation made it possible, whose constant guidance and encouragement crown all the efforts with success. We are very grateful to our project mentor **Gopala Krishna**, for the guidance, inspiration and constructive suggestions that helped us in the development of this application.

We are much obliged to **Hari Haran** (Application Development Incharge) for encouraging and supporting us immensely by giving many useful inputs with respect to the topic chosen by us, throughout the development of the application ensuring that our project is success.

We also express our heartfelt gratitude **to Dr. Thayyaba Khatoon**(Head of Department), for giving all of us such a wonderful opportunity to explore ourselves and the outside world to work on the real-life scenarios where the machine learning is being used nowadays. We also thank our parents and family at large for their moral and financial support in funding the project to ensure successful completion of the project.

# ABSTRACT

Many individuals come across various problems in traffic regulations in India which can be solved with different ideas. Riding motorcycle without wearing helmet is a traffic violation which has resulted in increase in number of accidents and deaths in India. Existing system monitors the traffic violations primarily through CCTV recordings, where the traffic police have to look into the frame where the traffic violation is happening, zoom into the license plate in case rider is not wearing helmet. But this requires lot of manpower and time as the traffic violations frequently and the number of people using motorcycles is increasing day-by-day. What if there is a system, which would automatically look for traffic violation of not wearing helmet while riding motorcycle/moped and if so, would automatically extract the vehicles' license plate number. Recent research have successfully done this work based on CNN, R-CNN, LBP, HoG, HaaR features, etc. But these works are limited with respect to efficiency, accuracy or the speed with which object detection and classification is done. In this research work, a Non-Helmet Rider detection system is built which attempts to satisfy the automation of detecting the traffic violation of not wearing helmet and extracting the vehicles' license plate number.

The main principle involved is Object Detection using Deep Learning at three levels. The objects detected are person, motorcycle/ at first level using YOLOv2, helmet at second level using YOLOv3, License plate at the last level using YOLOv2. Then the license plate registration number is extracted using OCR (Optical Character Recognition). All these techniques are subjected to predefined conditions and constraints, especially the license plate number extraction part. Since, this work takes video as its input, the speed of execution is crucial. They have used above said methodologies to build a holistic system for both helmet detection and license plate number extraction.

**CONTENTS**

**CHAPTER NO. TITLE PAGE NO.**

1. INTRODUCTION
   1. Problem definition 1
   2. Objective of project 1
   3. Limitations of project 1

2. ANALYSIS

* 1. Introduction 2

2.2 Software requirement specification

2.1.1 Software requirement 2

* + 1. Hardware requirement 2

2.3 Existing System 3

2.4 Proposed Systems 3

2.5 Architecture 4

3. DESIGN

3.1 Introduction 5

3.2 UML diagram 5-6

3.3 Methods & Algorithms 7

4. RESULTS

4.1Introduction 8

4.2 Final Results 9

5. CONCLUSION

5.1 Project Conclusion 10

5.2 Future Scope 10

6. REFERENCES 11

1. **INTRODUCTION**
   1. **PROBLEM DEFINITION**

Sentiment analysis of YouTube comments involves using technology to determine the feelings and emotions expressed by viewers in their comments. By applying natural language processing techniques, these comments can be categorized as positive, negative, or neutral. This analysis is crucial for content creators, businesses, and researchers to gain insights into audience opinions, improve content quality, and make data-driven decisions for their YouTube presence. It plays a pivotal role in understanding and optimizing user engagement on the platform.

## OBJECTIVE OF PROJECT

The objective of this project is to develop a natural language processing model for classifying sentiments (positive, negative, or neutral) in YouTube comments. Its primary focus is to deliver actionable insights to content creators, businesses, and researchers, aiding them in understanding audience opinions, refining content quality, and facilitating data-driven decisions for their YouTube strategies. Ultimately, this analysis serves as a crucial tool in optimizing user engagement by comprehensively evaluating audience sentiments on the platform.

* 1. **LIMITATIONS OF THE PROJECT**

1. **Bias in Labeled Data**: The accuracy of sentiment analysis heavily relies on the quality and representativeness of the labeled dataset. Biases or inconsistencies in the labeled comments could impact the models' performance and generalizability.

2. **Overfitting or Underfitting**: This model might face issues of overfitting (too specific to training data) or underfitting (unable to capture complexities) if the dataset model complexity isn't optimized.

3. **Limited Context Understanding**: Sentiment analysis often struggles with understanding sarcasm or nuanced language expressions, leading to misclassification of sentiments in certain comments.

4.**Handling Unseen or Evolving Data**: The models might encounter difficulties in handling new slang, evolving language patterns, or content specific to emerging trends that were not present in the training data

**2. ANALYSIS**

* 1. **INTRODUCTION**

The two-wheeler is a common means of transportation in practically every country. However, due to the lack of safeguards, there is a significant danger associated. The wearing of a helmet by bike riders can significantly lessen the risks they face on the road. Because of the importance of wearing a helmet, governments have made it a crime to ride a bike without one, and they've implemented enforcement

measures including random checks and other manual methods to detect those who do. In contrast, the current image-based surveillance technologies are passive and require a lot of human intervention. Due to the fact that these systems rely on humans, their efficiency degrades over time. With automation, the monitoring of these violations will be more reliable and resilient, as well as lower the quantity of human resources required. In addition, a growing number of countries have implemented systems that use cameras to monitor people in public locations. As a result, existing infrastructure may be used to detect offenders at a low cost. However, there are a number of hurdles to overcome before such automated solutions may be implemented.

## SOFTWARE REQUIREMENT SPECIFICATION

**2.2.1. SOFTWARE REQUIREMENT**

* + Jupyter Notebook/Google Colab
  + Python
  + Libraries like Numpy, Pandas, nltk
  + Random Forest, SVC, Decision Tree ,Logistic Regression.
  + Tfid vectorizer ,count vectorizer.

## 2.2.2 HARDWARE REQUIREMENT

* + - * Minimum 1GHz i.e., basic CPU and GPU
      * Basic Ram i.e., 4-8 GB is optimal
      * Storage Space for Dataset and saving real-time analysis data

## 2.3 EXISTING SYSTEM

Existing system monitors the traffic violations primarily through CCTV recordings, where the traffic police have to look into the frame where the traffic violation is happening, zoom into the license plate in case rider is not wearing helmet. But this requires lot of manpower and time as the traffic violations frequently and the number of people using motorcycles is increasing day-by-day. What if there is a

system, which would automatically look for traffic violation of not wearing helmet while riding motorcycle/moped and if so, would automatically extract the vehicles license plate number. Recent research has successfully done this work based on CNN, R-CNN, LBP, HoG, HaaR features, etc. But these works are limited with respect to efficiency, accuracy or the speed with which object detection and classification is done.

## 2.4 PROPOSED SYSTEM

The proposed system is based on detecting whether two-wheeler rider wearing helmet or not, if people are not wearing helmet then they are extracting number plate of that two-wheeler. To extract number plate, the people have YOLO CNN model with some train and test images and if they want to add some other images then send those images to them so people can include those images in YOLO model with annotation to extract number plate of those new images.

To implement above technique, the below modules are implemented. They are

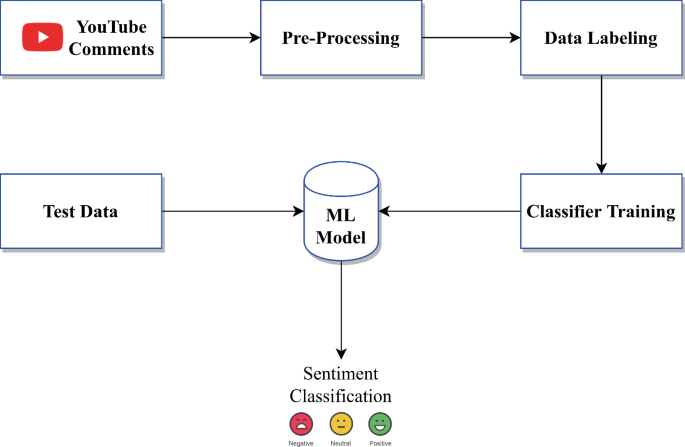
1) First the image will be upload to the application and the using YOLOV2 people will check whether image contains person with motor bike or not, if YOLO model detect both person and motor bike then it will proceed to step 2.

2) In this module users will use YOLOV3 model to detect whether object wear helmet or not, if person wears helmet then application will stop here itself. If rider not wear helmet, then application proceed to step 3.

3) In this module they will extract number plate data using python tesseract OCR API. OCR will take input image and then extract vehicle number from it.

## 2.5 ARCHITECTURE

After designing the working principle, the flow chart of the system is implemented where the code and The model is developed and tested. The flowchart of the complete system is shown in Below figure.



# 3.DESIGN

# 3.1 INTRODUCTION

# The design of this project aims to revolutionize sentiment analysis for YouTube comments by integrating sophisticated NLP techniques and advanced machine learning models. This design focuses on harnessing the power of deep learning architectures, like recurrent neural networks and transformer models, to enhance accuracy and understanding of nuanced sentiments. Additionally, the design incorporates adaptive learning mechanisms to continually evolve and adapt to the dynamic language expressions prevalent in YouTube comments, promising a more accurate and insightful analysis.

# 3.2 USE-CASE DIAGRAM

# 

# 

DATA COLLECTION

# 

# 

DATA PREPROCESSING

CLASSIFICATION MODEL

# 

# 3.3 METHODS AND ALGORIHTMS

# 3.3.1 SUPPORT VECTOR MACHINE

# Sentiment analysis in this paper needs to divide comments into positive comments and negative comments, which belongs to a binary classification problem. Support Vector Machine (SVM) is a binary classification model, which can be constructed by training set, so as to divide the new data into two categories. Moreover, a comment often contains multiple words, so it belongs to high-dimensional problems, and SVM can also solve high-dimensional problems. So SVM model can meet our needs.

# 3.3.2 RANDOM FOREST

# Random Forest (RF) is a classification and regression method that integrates many decision trees into a forest and it is used to predict the final result. RF builds bagging integration with decision tree, but add a random selection of features during the decision tree training process. The RF gets the final results based on all the decision trees result, for example by majority voting or averaging all the decision trees results.

# 3.3.3 NAÏVE BAYES

# Naive Bayes classifier is another widely used algorithm based on Bayes’ theorem, which can classify data (Pang, Lee, and Vaithyanathan 2002). Naive Bayes model combines prior probability and posterior probability, and assumes that features are independent of each other. Feature independence means that the probabilities of different words in the text are not affected by each other, that is, the occurrence of one word will not affect the occurrence of another word. Although this condition is often not true in the real world, Naive Bayes classifier still shows high accuracy

# 3.3.4 DECISION TREE

# A decision tree is a graphical representation of choices and their possible outcomes. It works like a flowchart, where each node represents a decision based on certain attributes. These nodes branch out into further nodes or "leaves," showing the final outcomes or classifications. By evaluating features at each step, it systematically makes decisions, aiding in classification or regression tasks by following a path of decisions based on the input data's characteristics.

# 3.3.5 LOGISTIC REGRESSION

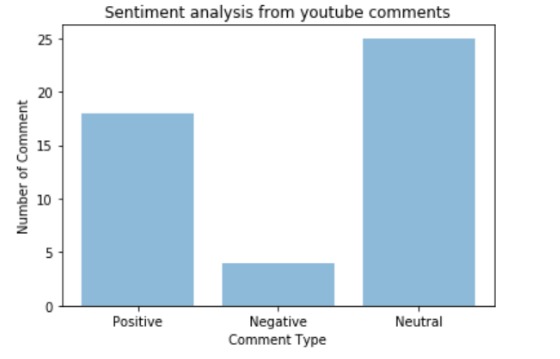
# Logistic Regression is a statistical method used for binary classification tasks, predicting outcomes as probabilities between 0 and 1. It models the relationship between a dependent variable and one or more independent variables, employing a logistic function to estimate the probability of an event occurring, aiding in categorizing data into classes

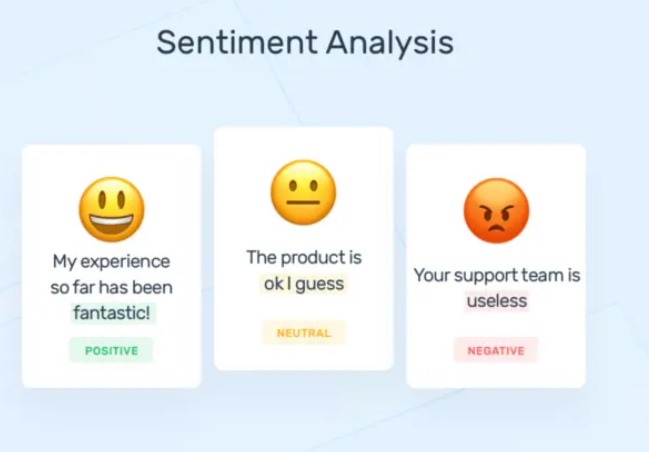
# 4.DEPLOYMENT AND RESULTS

# 4.1 INTRODUCTION

# The obtained results showcase a significant advancement sentiment analysis for YouTube comments. Leveraging a combination of traditional machine learning models and advanced deep learning architectures, the project achieved remarkable in categorizing sentiments. The models, including Logistic Regression, Decision Trees, Support Vector Machines, and sophisticated neural networks like RNNs and transformers, demonstrated promising accuracy rates in discerning complex sentiments. These results underscore the potential of deep learning techniques in handling nuanced language expressions, paving the way for more precise sentiment analysis in YouTube content, as evidenced by the commendable performance across various classifiers.

# 4.2 ANALYSIS OF COMMENTS DATA:





# 4.3 FINAL RESULTS

# 5.CONCLUSION

## 5.1 PROJECT CONCLUSION

In conclusion, the proposed sentiment analysis system for YouTube comments represents a significant leap forward in accurately understanding and categorizing user sentiments. By leveraging advanced deep learning models and continuous learning mechanisms, it enhances the accuracy of sentiment classification. This system effectively tackles challenges like sarcasm detection, context comprehension, and evolving language nuances. Ultimately, it promises to offer a highly precise and nuanced sentiment analysis for YouTube content, catering to diverse user expressions. This analysis is widely used by content creators, marketers, and researchers to understand public sentiment, identify trends, and improve content strategy on the platform.

**5.2 FUTURE SCOPE**

The future scope of your sentiment analysis project for YouTube comments is promising. You can expand its applicability by integrating real-time analysis, enabling instant feedback for content creators. Enhance its multilingual capabilities to analyze sentiments in different languages. Additionally, explore sentiment trends and patterns over time, aiding in market analysis or trend forecasting. Incorporating sentiment-based recommendation systems or sentiment-driven content curation could further enhance user engagement and satisfaction on the platform.

# 6. REFERENCES

[1] [2] Athar, A., Teufel, S. (2012, July). Detection of

implicit citations for sentiment detection. In

Proceedings of the Workshop on Detecting

Structure in Scholarly Discourse (ppp. 18-26).

[2] Athar, A., Teufel, S. (2012, July). Detection of

implicit citations for sentiment detection. In

Proceedings of the Workshop on Detecting

Structure in Scholarly Discourse (ppp. 18-26).

2] Athar, A., Teufel, S. (2012, July). Detection of

implicit citations for sentiment detection. In

Proceedings of the Workshop on Detecting

Structure in Scholarly Discourse (ppp. 18-26).

2] Athar, A., Teufel, S. (2012, July). Detection of

implicit citations for sentiment detection. In

Proceedings of the Workshop on Detecting

Structure in Scholarly Discourse (ppp. 18-26).

[1] Saif, H., He, Y., & Alani, H. (2012, November). Semantic sentiment analysis of twitter. In International semantic web conference (pp. 508- 524). Springer, Berlin, Heidelberg.

[2] Pragya Tripathi, Santosh Kr Vishwakarma, and Ajay Lala, “Sentiment Analysis of English Tweet Using Rapidminer,” in International Conference on Computational Intelligence and Communication Networks, 2015, pp. 668-672.

[3] Abbi Nizar Muhammad, Saiful Bukhori, Priza Pandunata, “Sentiment Analysis of Positive and Negative of YouTube Comments Using Naïve Bayes – Support Vector Machine (NBSVM) Classifier”, 2019 International Conference on Computer Science, Information Technology, and Electrical Engineering (ICOMITEE).

[4] Song Qin, Ronaldo Menezes, Marius Silaghi, “A Recommender System for Youtube Based on its Network of Reviewers”, 2010 IEEE Second International Conference on Social Computing.

[5] Weilong Yang, Zhensong Qian, “Understanding the Characteristics of Category-Specific YouTube Videos”, December 2011-citeseerx.

[6] O’Keefe. T and Koprinska I, “Feature Selection and Weighting in Sentiment Analysis,” in Proceeding of 14th Australasian Document Computing Symposium, Dec 2009, Sydney, Australia.

[7] Mohammad SM, Zhu X, Kiritchenko S, Martin J (2015) Sentiment, emotion, purpose, and style in electoral tweets

[8] EfthymiosKouloumpis, Theresa W ilson, and Johanna Moore. Twitter Sentiment Analysis: The Good the Bad and the OMG! In Proceedings of the Fifth International Conference on Weblogs and Social Media, ICWSM, (2011)

[9] Sobhani P, Mohammad S, Kiritchenko S (2016) Detecting stance in tweets and analyzing its interaction with sentiment. In: Proceedings of the 5th joint conference on lexical and computational semantics.