Sentiment Analysis of Youtube Live Chats using Machine Learning Methods

Aniket Malik
IIIT Delhi
aniket21231@iiitd.ac.in

Piyush Gautum
IIIT Delhi
piyush21549@iiitd.ac.in

Mayank Jha
IIIT Delhi
mayank20521@iiitd.ac.in

Abstract

In today's digital age, live streaming and real-time chat features have become an integral part of online entertainment. However, the ease of anonymous communication in live chats has given rise to various forms of toxicity, including hate speech and harassment, creating a hostile environment that adversely affects user experiences. While platforms often rate video content by age, live chat content remains largely unregulated.

This research is motivated by the pressing need to address this issue. Toxicity in live chats not only harms individuals but also tarnishes the reputation and community health of streaming platforms. Some platforms have taken steps to employ content moderation, but there is room for improvement. More refined machine learning models, trained on extensive chat data, can help detect nuanced forms of toxicity that elude traditional filters.

By enhancing toxicity detection algorithms and implementing real-time monitoring, streaming platforms can cultivate a safer and more inclusive online community for users of all ages. This proactive approach can mitigate the adverse effects of toxic behavior, fostering a more positive online environment.

1. Introduction

YouTube Live chats have transformed the way viewers engage with content creators in real time. These live chats provide an interactive space where viewers can communicate with each other and the streamer, sharing thoughts, questions, and reactions as events unfold. This dynamic feature has significantly contributed to the popularity of live streaming, making it a central aspect of online entertainment, educational webinars, and more.

However, the open and dynamic nature of YouTube Live chats also brings forth a critical concern - the prevalence of toxic behavior and harmful interactions within these spaces. Toxicity in live chats can manifest as hate speech, offensive comments, harassment, and other forms of disruptive communication. This not only creates a hostile environment but also detracts from the overall experience for users, potentially leading to serious emotional and psychological consequences.

Recognizing the gravity of this issue, our research focuses on Sentiment Analysis of YouTube Live Chats using Machine Learning Methods, with the primary objective of classifying chat messages as either safe or toxic. By doing so, we aim to foster a healthy and safe environment for live chat interactions among viewers and between viewers and streamers alike.

This problem holds immense significance, extending its use cases beyond the realm of entertainment. It can be applied to the moderation of online classroom chats, ensuring that students have a productive and respectful learning environment. Moreover, it can be utilized in various other scenarios where real-time chat interactions occur, ranging from live Q&A

sessions to community forums, enhancing the quality of online interactions and fostering a more positive and inclusive digital space.
In this paper, we will delve into the methodologies and machine learning techniques used to develop a model capable of accurately detecting and classifying toxic content in YouTube Live chats. We will explore the challenges and opportunities in this domain, with the ultimate goal of contributing to a safer and more welcoming online community**.**

2. Literature Survey

We look at two research papers that try to perform sentiment analysis or text mining. Following papers helped us better understand the use and difficulties that could be incurred while performing the analysis.

2.1 Surjuse, V., Dharne, A., &Lade, H. (2019). SENTIMENT ANALYSIS OF CHAT BASED APPLICATION USING R. Journal of Engineering and Technology Innovation Research.

https://www.jetir.org/papers/JETIR1903556.pdf

This paper aims to learn about the sentiments of an individual involved in the text conversation.
This task has several complexities:-
-Informal language and short forms
-Colloquialisms and slags.
The research is vital for several reasons:-
-Understand the effect of social media
-Study rising anxiety and depression due to social media usage.
-Study and curb the trends of cybercrime and online bullying
Therefore, this research has implications in fields like management and social sciences, not just computer science.
One approach to sentiment analysis is,
Lexical approach = It uses literal keywords like 'happy,' 'sad,' 'afraid,' etc and their dictionary synonyms. It is simple and naive but does not understand negation. It also does not understand the subtext and relies on surface keywords. Therefore, it is prone to error.
Machine Learning-based approach uses three stages: Data collection, Pre-processing, Training data, Classification and plotting results.

2.2 Chouhan,A.,Halgekar,A.,Rao,A.,Khankhoje, D. & Narvekar,M.(2021).Sentiment Analysis of Twitch.tv Livestream Messages using Machine Learning Methods.Dwarkadas J. Sanghvi College of Engineering.

https://ieeexplore.ieee.org/document/9616932

This paper offers a methodology for sentiment analysis with ML models on live stream messages.
The task presented several challenges:
A large number of messages were involved.
Streamer-specific slang and emoticons made the language context-dependent.
The language contained numerous abbreviations, repetitions, and deliberate grammatical mistakes.
The use of emojis added complexity to the problem.
The following workflow was employed:
-They had a labeled dataset.
-Pre-processing was performed, and the dataset was split into three sets for training, validation, and testing (60:20:20).
-The pre-processing pipeline comprised three steps: tokenization, removal of stop words, and lemmatization.
-Stop words, which are common words with little to no meaning in text analysis, were removed.
-Data was then split.
-Subsequently, the text was converted into vector form using Count Vectorization and TF-IDF Vectorization.
Different ML models were used and their accuracies were compared.
Support Vector Classifier 70.4%
Logistic Regression 69.2%
Decision Tree Classifier 67.2%
Random Forest Classifier 66.4%
Multinomial Naïve Bayes 65.8%
SVM produced the best accuracy of SVM 70.4% suggesting that it might be the best model for us.

3. Dataset

The following dataset was used-
uetchy. (2021). Sensai Dataset. Kaggle. URL: https://www.kaggle.com/datasets/uetchy/sensai/data
As the dataset was divided into a number of parquet files,we combined them into a single CSV file.
The dataset employed in this research study comprised a substantial dataset, consisting of a total of 7,700,072 rows. These rows were structured around two primary columns, namely 'body' and 'label.'Label were of three types-

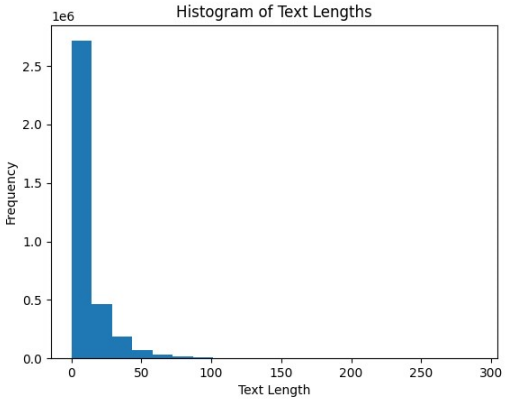

Fig.1.Histogram for Frequency of Length of Chats

-Hidden/Deleted: For flagged(toxic) chats
-Nonflagged: For non-flagged chats

Chats contained both ASCII and non-ASCII characters along with special characters. The following techniques were used in Data Preprocessing-
Removing Duplicates and NAN/empty strings-
Duplicates were removed from the dataset and the NAN/empty strings were deleted thus reducing the size from 7,700,072 to 3,508,663 rows.
Then we used the following techniques-
a)Lowercasing

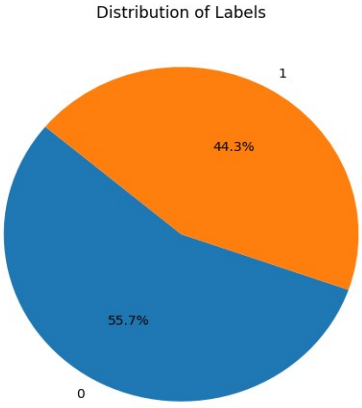

Fig.1. Pie Chart depicting Distribution of Labels

b)Removing Non-ascii and special characters
c)Tokenization

d)Normalizing Contractions
We also label encoded the labels-
-Hidden/Deleted: 1
-Nonflagged: 0
The dataset is well balanced as it contains 44.3% label 1 and 55.7% label 2 indicating the dataset is not biased.
We also analysed the frequency of the length of chats to help understand the feasibility of pre-processing and training on a large dataset.

4. Methodology and Model details

Initially, to perform preprocessing and data analysis, we used several libraries in Python including - numpy, pandas,sklearn, sea born, matplot.lib, contractions.
Preprocessing is vital in text sentiment analysis to cleanse and structure text data, removing noise and irrelevant information, ensuring accurate sentiment classification and meaningful insights.
Data preprocessing-
a) Removing Duplicates-Removed duplicates rows for better accuracy and preventing biasing.
b)Lowercasing: Converted all characters into lowercase.For eg- otherwise man and Man would be treated differently by the model.
c)Tokenization: Splitted the large sentences into smaller chunks which will help in preprocessing of the data more efficiently and reasonably using word_tokenize from NLTK library
d)Normalizing Contractions-Contractions were expanded during data processing; for instance, "don't" was transformed into "do not" to facilitate comprehensive text analysis
e)Number of Samples split: Samples from the column of the label has been divided in half -
-Hidden/Deleted: 1
-Nonflagged: 0
Feature Extraction-Feature extraction is the process of transforming raw data, often high-dimensional or unstructured, into a reduced set of meaningful and informative features that can be used for analysis, modelling, or machine learning tasks.We implemented this using -
TFIDF Vectorizer-The TFIDF Vectorizer was employed to convert textual sentences into numerical vectors, serving as a standardization and model representation technique for building various models, such as logistic regression and SVM.
Postprocessing, the dataset was reduced in size and better suited to ML models.
The following ML models were used-
We used an 80:20 (train: test) data split for our models.
Logistic Regression-Logistic regression is a supervised machine learning algorithm mainly used for classification tasks where the goal is to predict the probability that an instance belongs to a given class. It is used for classification algorithms its name is logistic regression. it's referred to as regression because it takes the output of the linear regression function as input and uses a sigmoid function to estimate the probability for the given class.
Max_Iterations:10,000 Random State:42

5. Result and Analysis

We obtained the following results after performing logistic regression-

Accuracy	0.6511450936467289
Precision	0.7032274331820474
Recall	0.3678058508499654
F1 Score	0.48299370011890147

Table 1.Logistic Regression Results

The analysis of the logistic regression results indicates the following:
-Accuracy (0.6511): This metric represents the overall correctness of the model's predictions. An accuracy of approximately 65.11% suggests that the model correctly classified around 65.11% of the data points, which may be considered moderate but not exceptionally high.
-Precision (0.7032): Precision measures the proportion of true positive predictions among all positive predictions made by the model. In this case, a precision of approximately 70.32% indicates that

when the model predicts a positive outcome, it is correct about 70.32% of the time.
-Recall (0.3678): Recall (also known as sensitivity) measures the proportion of true positive predictions among all actual positive instances. A recall of about 36.78% suggests that the model captured only 36.78% of all actual positive instances, indicating room for improvement in identifying positive cases.
-F1 Score (0.4830): The F1 score is the harmonic mean of precision and recall, providing a balanced measure of a model's performance. An F1 score of approximately 48.30% indicates a trade-off between precision and recall, showing that there may be a balance between false positives and false negatives.

Label	Precision	Recall	F1 Score	Support
0	0.64	0.88	0.74	390838
1	0.70	0.37	0.48	310895

Table 2.Classification Report

The classification report provides a more detailed breakdown of the model's performance for each class (0 and 1), along with macro and weighted averages:
-Accuracy: The overall accuracy of the model is 65%, as previously mentioned.
-Macro Avg: The macro-average calculates the average performance across both classes. In this case, it indicates that the model's average precision, recall, and F1-score across the two classes are 0.67, 0.62, and 0.61, respectively.
-Weighted Avg: The weighted average considers the class imbalance, giving more weight to the class with a larger number of instances. The weighted average precision, recall, and F1-score are 0.67, 0.65, and 0.62, respectively.
Analysis:
-For Class 0 (Negative Class), the model shows relatively good performance with a high recall (0.88) but slightly lower precision (0.64). This indicates that the model correctly identifies the majority of negative cases but may have some false positives.
-For Class 1 (Positive Class), the model's precision (0.70) is decent, indicating that when it predicts a positive case, it's often correct. However, the recall (0.37) is lower, suggesting that the model misses a significant portion of actual positive cases.
The overall F1-score for the model is 0.62, which is a balanced measure of precision and recall. It indicates that there is room for improvement, especially in correctly identifying positive cases (Class 1).

5. Conclusion

Based on the results,we can conclude that the logistic regression model achieved moderate accuracy and precision, but it struggled with recall. Therefore logistic regression is unlikely the best fit model for our case, other models like SVM,Decision Trees are likely to perform better.

`