Computer Science Capstone

Western Governors University

Cade Fisher

October 6, 2024

Part A: Letter of Transmittal

October 6, 2024

WGU Evaluators
Western Governors University
4001 S 700 E Millcreek UT, 84107

Dear WGU,

I am pleased to present a proposal for a new application designed to address the growing issue of toxic comments on our platforms. As our user base continues to expand, the challenge of managing harmful content has become increasingly significant. This application leverages machine learning to analyze YouTube comments for various types of toxicity, providing us with essential insights to foster a safer online community.

The primary purpose of this application is to identify and categorize toxic comments, such as hate speech, abuse, and threats. By implementing this tool, we will enhance our moderation process, enabling us to address harmful content more efficiently. The application not only automates the identification of toxic comments but also provides detailed reports, allowing us to understand the nature and frequency of these comments over time. This information is vital for improving our community engagement strategies and creating a more positive environment for our users.

In terms of implementation, the project is estimated to take around three months and will involve data collection, model training, and application development. The overall budget will cover development costs, ongoing maintenance, and any necessary software licenses. We have also taken ethical considerations into account, ensuring that user data is handled responsibly and transparently.

I believe my background in data science and experience with machine learning make me well-suited to lead this project. Together, we can create a powerful tool that not only protects our community but also enhances user satisfaction and trust.

Thank you for considering this proposal. I am looking forward to discussing it with you further and exploring how we can make our online spaces safer for everyone.

Sincerely,

Cade C Fisher

Cade C Fisher

Part B: Project Proposal Plan

Project Summary

Problem Statement

Toxic comments on YouTube videos are an ongoing issue that affects both content creators and their audiences. Negative comments can create a hostile environment, discourage engagement, and harm the mental well-being of creators. To address this problem, we need to develop a tool that helps creators manage and analyze the toxicity of comments on their videos effectively.

Client Overview and Needs

Our primary clients will be YouTube content creators, ranging from influencers to educators. They are looking for a solution that allows them to understand the nature of comments being posted on their videos. By analyzing comment toxicity, creators can identify harmful interactions, improve community engagement, and foster a positive atmosphere for their viewers. They need a user-friendly tool that provides insights without requiring extensive technical expertise.

Deliverables

1. **Comment Analysis**
    1.1. The core functionality of the tool will analyze comments from a provided YouTube video link. It will assess the comments for various types of toxicity, including abusive, threats, provocative, obscene, hate speech, racism, nationalist, sexism, religious hate, radicalism, and homophobic language, providing creators with insights into the sentiment of their audience.
2. **Visual Graphs**
    2.1. The application will include two visual graphs to illustrate trends in comment toxicity.
        2.1.1. Quantitative Bar Graph of Toxic Comments: This graph will display the number of comments in each specific toxicity category.
        2.1.2. Percentage Graph: This graph will present the percentages of toxic comments and indicate which categories they belong to.
3. **User-Friendly Website Interface**
    3.1. We will develop a website that offers a simple and intuitive interface for users. This site will allow content creators to easily input video links, view analysis results, and explore visual graphs without needing technical expertise.

Justification of Benefits

The implementation of this tool will provide several key benefits for our clients. First, it will enable effective comment management, allowing creators to quickly identify and address toxic comments, thereby fostering a safer online environment for their audiences. Additionally, the insightful trend analysis offered by the visual graphs will help creators track changes in comment toxicity over time, facilitating informed decisions about their content and community engagement strategies. Furthermore, the user-friendly interface will ensure that creators can navigate the tool with ease, allowing them to spend more time on content creation and less on comment moderation. In summary, the Toxic Comment Analysis Tool will empower YouTube content creators to proactively manage their comment sections, resulting in a healthier and more engaged online community. By providing clear insights into various types of toxicity, we will help them cultivate a positive atmosphere for their viewers.

Data Summary

Source of Raw Data

The raw data for this project will be sourced from a dataset available on Kaggle, specifically from the link: https://www.kaggle.com/datasets/reihanenamdari/youtube-toxicity-data We will collect the data by downloading it in CSV format and then using programming to extract and manipulate the relevant information.

Data Processing and Management

Throughout the application development lifecycle, the data will be carefully processed and managed. Initially, the data will undergo a cleaning phase to remove any irrelevant or redundant information, ensuring that it is suitable for analysis. After cleaning, the data will be formatted into a dataset compatible with machine learning algorithms, allowing for effective training and evaluation.

Justification of Data Suitability

This dataset meets the project's needs as it consists of 1,000 comments that have been categorized into specific toxicity categories. This categorization will enable the machine learning model to learn from key words and phrases, thereby enhancing its ability to analyze and classify new comments accurately. Any data anomalies, such as outliers or incomplete entries, will be addressed during the cleaning process to ensure the integrity and reliability of the dataset.

Ethical and Legal Concerns

Regarding ethical and legal concerns, there are no issues with this data, as it is publicly available. This accessibility ensures that we can use the data without infringing on any copyrights or privacy rights. Overall, the data will serve as a robust foundation for developing the Toxic Comment Analysis Tool.

Implementation

Industry standard methodology

We will utilize the Agile methodology for this project, which emphasizes iterative development, collaboration, and flexibility. Agile allows for regular feedback from stakeholders and the ability to adapt to changes throughout the project lifecycle. This methodology will help ensure that we can incorporate user insights and make necessary adjustments to the tool as we progress.

Project Implementation Plan

The implementation of this project will follow a structured plan that includes the following key phases. By following this implementation plan, we will ensure a systematic approach to developing the Toxic Comment Analysis Tool, resulting in a reliable and user-friendly application that meets the needs of YouTube content creators.

1. Data Gathering:
    1.1. Download the YouTube toxicity dataset from Kaggle.
    1.2. Extract relevant data from the CSV file for analysis.
2. Data Cleaning:
    2.1. Remove any irrelevant or redundant entries from the dataset.
    2.2. Address data anomalies, such as outliers or incomplete data, to ensure the dataset is reliable.
    2.3. Format the comments in a way that makes them easier for the machine learning model to read and process. This may include standardizing text (e.g., lowercasing, removing special characters)
3. Machine Learning Model Development:
    3.1. Develop a neural network-based machine learning model to analyze comment toxicity.
    3.2. Train the model using the cleaned and formatted dataset, focusing on key features that indicate toxicity.
    3.3. Validate the model's performance and adjust as needed based on testing results.
4. Website Development:
    4.1. Create a user-friendly website interface for content creators to input video links and view analysis results.
    4.2. Implement the comment analysis feature and integrate the machine learning model with the website.
5. Visual Graphs Creation:
    5.1. Develop the quantitative bar graph and percentage graph to visualize comment toxicity trends.

5.2. Ensure that the graphs are easily interpretable and provide actionable insights for users

6. Testing and Quality Assurance:
    6.1. Conduct thorough testing of the application to identify and fix any bugs or usability issues

Project Timeline

Milestone or Deliverable	Duration	Project Start Date	Anticipated End Date
Data Gathering	1	September 30, 2024	October 1, 2024
Data Cleaning and Formatting	2	October 1, 2024	October 3, 2024
Machine Learning Model Development	5	October 3, 2024	October 8, 2024
Website Development	6	October 8, 2024	October 14, 2024
Visual Graphs Creation	5	October 14, 2024	October 19, 2024
Testing and Quality Assurance	3	October 19, 2024	October 22, 2024
Deployment and Maintenance	2	October 22, 2024	October 24, 2024

Evaluation Plan

Verification Methods at Each Stage of Development

1. Data Gathering
    1.1. After downloading the dataset, we will perform a review to ensure the data is complete and correctly formatted. This will include checking for the presence of all expected columns and confirming that the dataset contains 1,000 comments as stated.
2. Data Cleaning and Formatting
    2.1. We will conduct tests to confirm that the cleaning process effectively removes irrelevant entries and standardizes the text format. This will include sample checks of the cleaned data to ensure no critical information is lost and that the comments are consistently formatted.
3. Machine Learning Model Development
    3.1. We will utilize neural networks for our machine learning model and manually test a small sample of comments to verify that the model accurately classifies them. This manual testing will help confirm the model's functionality before moving on to larger datasets.
4. Website Development
    4.1. We will manually perform a quality assurance (QA) test on the website to ensure that all features function as intended. This will include testing the input fields, output displays, and overall user experience.
5. Visual Graphs Creation

5.1. We will validate the accuracy of the visual graphs by cross-referencing them with the raw data. This will ensure that the graphs correctly represent the underlying data trends and categories.

Validation Method Upon Completion of the Project

Upon completing the project, we will use a clear validation method that includes several important steps. First, we will conduct end-to-end testing to check that the entire system works properly, from data input to analysis output. This will involve simulating real situations where users enter YouTube video links and get the analysis results. Next, we will present the final product to key stakeholders in a review session, collecting their feedback to ensure the tool meets the original project requirements and user needs. Lastly, we will evaluate the performance of the machine learning model using key metrics like precision, recall, and F1-score to confirm its ability to accurately identify toxic comments. This thorough approach will ensure that the Toxic Comment Analysis Tool is reliable and ready for use.

Resources and Cost

Resource Type	**Item**	**Cost**	**Time**	**Total cost**
Hardware	Laptop	$1,500.00	1	$1,500.00
Software	Python	$-	0	$-
Software	Scikit-learn	$-	0	$-
Software	Flask	$-	0	$-
Software	PyCharm	$-	0	$-
Labor	Data Engineer	$40.00	40	$1,600.00
Labor	Web Developer	$30.00	40	$1,200.00
Labor	QA	$25.00	20	$500.00
Labor	Developer	$40.00	60	$2,400.00
Environment	Cloud Hosting	$300.00	1	$300.00
Total				**$7,200.00**

Part D: Post Implementation Report

Solution Summary

The project addressed the problem of toxic comments on YouTube, which can create a negative experience for both content creators and their audiences. With so many comments, it can be hard for creators to identify and manage harmful content effectively. To solve this issue, we developed the Toxic Comment Analysis Tool. This application uses machine learning to

analyze comments from YouTube videos. Users simply enter a video link, and the tool extracts the comments, categorizing them into different types of toxicity, such as hate speech or abuse.

The application not only detects toxic comments but also provides visual representations, including bar graphs and pie charts, to show the distribution of comment categories. Additionally, it generates a heatmap that visualizes the performance metrics of the classification model, giving creators insights into how well the tool performs. These visual tools help creators quickly see trends in their comments, enabling them to manage their sections more effectively. By using this tool, content creators can create a safer and more positive environment for their viewers.

Data Summary

Data Source and Collection

The raw data for the project was sourced from a publicly available dataset on Kaggle, specifically designed for analyzing toxic comments on social media platforms. The dataset contained a collection of comments categorized into various toxicity types, which allowed for effective training of the machine learning model. The data was collected by downloading the dataset as a CSV file, which served as the foundation for our analysis.

Data Processing and Management

Throughout the development of the application, the data was carefully processed and managed in several important stages. In the initial design phase, we recognized the need for a clean and structured dataset to train the machine learning model and planned to preprocess the text for better accuracy. During development, we removed punctuation, converted text to lowercase, and tokenized comments, making the data easier to analyze. We then split the dataset into training and testing sets to assess the model's performance. The processed data was used to train a multi-output classifier, fitting the model on the training data while checking its performance with the testing set. After deployment, we monitored the data for any issues, like outliers or inconsistencies. We also planned for regular updates and retraining to incorporate new comments and improve accuracy. This organized approach ensured that the data remained clean and effective, leading to better predictions and insights into toxic comments.

Machine Learning

Method Identification

The primary machine learning method employed in this project is the MultiOutputClassifier using an MLPClassifier. This method is designed to handle multi-label classification tasks, meaning it can identify multiple types of toxicity in a single comment, such as whether it is abusive, racist, or homophobic.

Development Process

The development of this method involved several steps. First, we preprocessed the text data by removing punctuation, converting it to lowercase, and tokenizing the comments. This preparation made the data suitable for analysis. Next, we used the TfidfVectorizer to transform the text into numerical format, allowing the machine learning model to process it. We then split the dataset into training and testing sets to train the classifier and evaluate its performance. The MLPClassifier was configured with a maximum of 1000 iterations and fitted using the training data. After training, we used the classifier to predict the categories of toxicity for new comments.

Justification for Selection and Development

The choice of the MultiOutputClassifier with MLPClassifier was based on the need to classify multiple types of toxicity in comments at the same time. This is important for understanding the different ways comments can be harmful. Neural networks, especially MLPs, are good at recognizing complex patterns in data, making them a great fit for text classification. This method not only boosts prediction accuracy but also provides deeper insights into toxic comments. Furthermore, our organized approach ensured that the data was clean and relevant, which helped improve the model's performance and reliability in real-world situations.

Validation

Method

To evaluate the performance of the MultiOutputClassifier using the MLPClassifier, we employed several key metrics: precision, recall, and F1-score. Precision measures the accuracy of the positive predictions, recall assesses the model's ability to find all relevant instances, and the F1-score provides a balance between precision and recall. By generating a classification report on the testing set, we obtained a comprehensive view of how well the model can classify toxic comments across multiple categories.

Future Validation and Monitoring

For future assessments, we plan to continuously monitor the model's performance by updating the dataset with new comments and retraining the model regularly. This ongoing evaluation will help maintain accuracy as language and online behavior evolve. Additionally, we intend to implement cross-validation techniques to ensure that our results are reliable and not due to overfitting. This systematic approach will enable us to adapt to changes in user-generated content and improve the model's effectiveness in identifying toxic comments.

Visualizations

The application features three unique visualizations that provide insights into the analysis of toxic comments. These visualizations are accessible from the results page that appears after a user inputs a YouTube video link.

- Bar Graph: This visualization displays the distribution of toxic comment categories, allowing users to see the number of comments falling into each toxicity type. It helps in understanding which categories are most prevalent in the analyzed comments.
- Pie Chart: The pie chart visualizes the percentage of toxic comments by category. This representation provides a clear view of the proportion of different types of toxicity, making it easy to grasp the overall composition of toxic comments.
- Heat Map: The heat map presents key performance metrics of the machine learning model, including precision, recall, and F1-score for each toxicity class. This visualization enables users to quickly assess the model's performance and identify areas for improvement.

User Guide

Steps to Execute and Use the Application

1) A zip file containing the application will be provided. Download this file to your computer.
2) Locate the downloaded zip file and unzip it to extract its contents.
3) You will need a Python IDE like PyCharm run the application.
    a) Ensure you have Python installed on your system.
4) Navigate to the directory where you unzipped the application.
5) Install the required libraries
6) Run the Application
7) Access the Application:

    a) Open a web browser and enter the following URL in the address bar: http://localhost:5000

    b) On the homepage, enter a valid YouTube video link (note: do not use a playlist link).

    c) Click the “Analyze” button.

8) View Results:

    a) After the analysis, you will see buttons to view individual comments, a bar graph, a pie chart, and a heat map.

    b) Click on each button to see the respective visualizations.

    c) There will be a link labeled “Analyze Another Video.” Click this link to return to the homepage and enter a new YouTube video link for analysis.