National College of Ireland

Sentiment Analysis of User comments for a YouTube Educational videos

MSc Research Project
Masters of Science in Data Analytics
(MSCDAD_A_JAN24I)

Sayali Jadhav
Student ID: x23201665

School of Computing
National College of Ireland

Supervisor: Prof. Vladimir Milosavljevic

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Sayali Jadhav
Student ID:	x23201665
Programme:	Masters of Science in Data Analytics (MSCDAD_A_JAN24I) **Year:** 2024-2025
Module:	MSC Research Project
Supervisor:	Prof. Vladimir Milosavljevic
Submission Due Date:	12/12/2024
Project Title:	Sentiment Analysis of User Comments for a YouTube Educational Videos
Word Count:	6119 **Page Count** 19

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Sayali Jadhav
Date:	12/12/2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Sentiment Analysis of User comments for a YouTube Educational videos

Sayali Jadhav
x23201665

Abstract

Over the years, online learning has outgrown and has become a critical component for educational area. Variety of instructors and video makers based on various online platforms like Udemy, Coursera and YouTube have transformed education by providing anytime on-click courses, which attracts a global audience. While there is extensive research already been conducted on analysis of sentiments over traditional classrooms and other online learning content, but we see that less focus has been given to the vast repository of YouTube's user comment. This gap emphasizes the need to understand the emotional components of the educational information offered through this medium. This study bridges the gap between content providers and viewers by applying sentiment analysis to analyse the YouTube user comments on educational content providing a thorough examination of YouTube comments. Approaches like TF-IDF and NRC lexicon are used in combination to identify the sentiment polarity from negative to positive (-1 to 1). Visualizations in form of polarity graphs, helps understand sentiment distributions and their implications. Our approach gives instructors meaningful insights by allowing them to understand the audience responses, identify the growth areas and increase engagement of learners. It provides the instructors with useful insights on learner emotions and encourages them to improve the online learning experience. Future research intends to improve the emotion analysis and address the current issues in individualized education.

Keywords: Sentiment Analysis, YouTube, Comments, Educational, Emotions, Learning, TF-IDF, Data pre-processing, NRC Lexicon, Polarity, Instructors

1 Introduction

The education industry has undergone a big shift as part of modern education world where people embrace online learning. Udemy, Coursera and YouTube are some platforms that have made all of this possible by making the learning materials accessible to different parts of the world where students get various educational content related to their search. These platforms have their drawbacks too as students are still able to learn through active classrooms but in different settings where they are not simply listening but are able to search for answers and give reviews. Also, there is major gap in understanding and communication between the instructors and the students. As in most cases, instructors are unaware from knowing if the students are viewing the clips for or how much they are consuming the knowledge from their material.

This research seeks to fill the gap by explaining sentiment analysis from the comment section of the videos by understanding the sentiments expressed in the comments.

Sentiment Analysis which is the subclass of Natural Language Processing, is defined as the computational study of opinions, emotions and sentiments in the writing. By determining whether the piece of review is neutral, positive or negative. Applying sentiment analysis on user reviews through comments from educational videos provide a systematic approach to understand the viewers experience and issues. This information can help the instructors to refine their teaching practices, modify the content as per learner’s expectations, and

eventually improving the overall quality of education offered through the online e-learning platforms.

Background:

The growth of e-learning platforms has underlined the importance of the user reviews and feedback in creating effective content. Also, sentiment analysis has been extensively researched in the context of online platforms and social media but its applicability to instructional e-learning content of YouTube is still underexplored.
Previous studies like those by Maw et al. (2024), Akila et al. (2022),
Nithyashree et al. (2020), have shown the effectiveness of the sentiment analysis in identifying learner's sentiments. However, this research generally focuses on social media and traditional approaches, creating a gap in understanding the emotional components of viewers engagement with YouTube's vast repository of educational videos.
This gap will not only bridge the distance between students and instructors as students will not only consume what is spoken in the video but also actively participate with the material via comments and get responses back. These comments mark valuable source of insights into students understanding level, their difficulty area and know their opinions. This research intends to bridge the gap between content creators and their audiences by analysing sentiments from the comment sections of YouTube educational videos using combination of Natural language processing and Data Mining approaches.

Motivation:

This research is motivated by the desire to enhance the quality and relevance of online educational content. As instructors too might face challenges in presenting content appropriate or as expected from audience with their different needs and variations. By analysing comments from the audience, educators can acquire a better understanding of student sentiments, allowing them to address their specific issues, identify areas of improvements, and create a more engaging learning online environment.
Also, sentiment analysis provides an efficient and scalable approach for processing large scale of unstructured text data, making it an ideal tool for evaluating and analysing YouTube comments. This work provides a systematic structure for evaluating and improving the YouTube educational content by categorizing the comments and defining the sentiment polarity either as positive, neutral or negative.

Research Question and Objectives:

This research is formulated by the central question as:
How effective is the application of sentiment analysis on comments of YouTube educational video in identifying the areas of development using Natural Language Processing and Data Mining techniques?

To address this question, the work establishes following research objectives:

1. Investigating the present state of work using sentiment analysis with focus on YouTube educational video comments section.

2. Designing a structured approach for extracting, preprocessing, and evaluating the comment textual data from YouTube educational videos.

3. Implementing sentiment analysis to identify emotions and analyse the polarity of sentiments from negative to positive using techniques like TF-IDF and NRC Lexicon.

4. Visualizing the end results to draw actionable insights through polarity graphs and other data visualization techniques.

Contribution:

The primary contribution of this research is the comprehensive analysis of comments of YouTube educational videos, which includes descriptive annotations of user inputs through feedbacks to help bridging the communication gap between students and instructors. This research provides practical insights about student sentiments, allowing educators to improve the effectiveness of their teaching content and relevance of their videos.
Additionally, the study introduces a systematic framework to analyse the comments, which altogether include data collection, data preprocessing, sentiment analysis and visualization of the results. This approach not only addresses the issues related to unstructured textual data but also provide scalable solutions for evaluating feedbacks on a larger scale.

Structure of the Paper:

The research paper is organized into following sections:
Section 1: Related Work
This research section elaborates and provide insights about the related or aligned existing research papers being worked in the area of our research i.e. sentiment analysis and sentiment analysis of comments from online videos. Also, it gives us a direction to our research enhancements we can approach from other related work.

Section 2: Research Methodology
This section of the report provides the step-to-step details from data extraction to visualization. It indicates the flow of the research steps carried out for visualizing the output.

Section 3: Design Specification
This section of the research provides the framework of associated requirements by underlining the implementation structure.

Section 4: Implementation
This section implements the designed framework derived through visualization outputs at end.

Section 5: Evaluation
This section evaluates the outcomes and provides the findings we received from the implemented framework. It explains the findings in brief with details

Section 6: Conclusion and Future Work
This section of the report summarises the research paper and provide the future actionable insights which can be carried through.

2 Related Work

Sentiment analysis has evolute rapidly which has driven research in variety of fields, including social media, education and consumer behavior. This literature review section gives a context about the current study by studying or examining the previous contributions to the field, identifying their limitations and explaining the requirement for new research. The reviewed research work is structured based on the methodologies, focusing on key findings, problems encountered and gaps.
This section will provide detail insights by reviewing existing technologies and topics including their applications. It will highlight gaps and flaws in existing research content.

Sentiment Analysis Applications

Sentiment Analysis of E-Learning platforms:

The diversity of feedbacks received as form of comments in the e-content and the importance of preprocessing steps like tokenization and noise removal was analysed by Singh and Tiwari in 2021. They emphasized on the YouTube comments to estimate the learner sentiment through it. They achieved 82% accuracy with their work. Despite notable contributions their methods faced issue to detect contextual polarity in remarks. This constraint emphasized the need for models which can handle linguistic variances, which current research work tries to address by using NRC emotion lexicon and data mining approaches.

Another research introduced a hybrid approach which combines Support vector Machine and Naive Bayes for doing sentiment analysis in educational videos. This research was proposed by Rajesh and Akila (2022) where they focused on identifying the challenges that are faced by the students to understand the content of the course. This hybrid model achieved 89% accuracy in classifying polarity but struggled with classifying nuanced sentiments. While their approach was effective but it lacks the real-time analysis and integration with visualization tools. While this research project intends to address by adding Natural Language Processing approach.

Sentiment Analysis of Social-Media:

A lexicon-based sentiment analysis framework for social media was created by Naresh Kumar and Uma (2021) where they highlighted the importance of context-aware models by achieving accuracy of 86%. Their research face limitations of managing domain-specific emotions which is part of generic lexicons. This research study builds on their findings by precisely analysing student sentiments using NRC Lexicon which has been adapted for educational content.

SentiDiff, a model for sentiment analysis of Twitter data which combines textual information with patterns was introduced by Wang et al. in 2019. Their findings emphasized the significance of sentiment context and diffusion in understanding user behavior which achieved 90% accuracy. While their research work focuses on social media, their work lay the base for applying sentiment diffusion approach to educational platforms, allowing for more comprehensive analysis of YouTube comments.

Emojis and Non-Traditional Inputs: Their Role in Sentiment Analysis:

The effects of emojis in sentiment analysis was investigated by Shiha and Ayvaz (2017) where they found out that they can have a significant impact on sentiment classification accuracy metrics. Their work revealed that inclusion of emoji sentiment mapping into sentiment analysis models improves and increase the user understanding by results upto 85%. Their research was largely focused on social media data and did not take educational online feedback into account. Our research incorporates the educational YouTube comments with sentiments and emojis analysis.

Other research which proves a base regarding educational content on YouTube's noting that the platform is good in delivering quality content. Scoping analysis was conducted in 2022 by Shoufan and Mohamed on the YouTube e-content. However, their research work limited a focus on user feedback analysis which is essential for improvement of the content. This work addresses the gap by the use of sentiment analysis on the YouTube comments section to provide instructors with actionable insights into student perception.

Comparative and Deep Learning Approaches

Sentiment Analysis using Comparative and Hybrid Models:

The significance of identifying sentiment diffusion patterns for accurate sentiment classification was done by Athindran et al. (2018) who examined the competing brands using hybrid sentiment analysis modelling. This concept can be used to track sentiment patterns and trends over the time in educational content, allowing content creators to monitor changes in student responses over feedbacks.

Research was conducted by merging machine learning models and sentiment lexicons which improve the accuracy to 87% of sentiment classification. This work was put forward by Rajeswari et al. (2020) where they used hybrid methodologies to compare customer sentiments. Their study underscores the requirement of hybrid frameworks, which the current research study implements.

Sentiment Analysis using Deep learning Approaches:

Deep learning approache was used by Maw et al in 2024 to analyse YouTube comments which achieved improved accuracy upto 92% by using complex model BiLSTM. Their research results revealed the robustness of deep learning in handling larger datasets and complex text structures. But they identified challenges with model interpretability and computational efficiency. To build on the prior work, this study applies data visualization approaches to improve the sentiment analysis results interpretability.

A BiLSTM model was developed by Xu et al. (2019) for Chinese text sentiment analysis. The importance of domain specific sentiment dictionaries in improving accuracy upto 88% of the model was underscored in this work. While effective in their context, the authors did not discuss the application of such similar methodologies with non-standardized or multilingual content, such as YouTube comments. This study help building on their findings by applying sentiment analysis to diverse languages and their patters from YouTube comment.

Limitations and Challenges of Sentiment Analysis Previous Work:

Despite earlier research work regarding sentiment analysis, some issues and limitations still remain. These include dealing with addressing many factors like computational inefficiencies, linguistic diversity and accurately interpretating mixed sentiments. Additionally, the use of sentiment analysis on YouTube educational videos remains underexplored, with few research study concentrating on adapting insights to improve teaching strategies by addressing teaching content online.

Research: Justification and Contribution

In conclusion, this existing research reviews identifies the limits of existing sentiment analysis models. The issues it includes are polarity, linguistic variety, and real time processing. It underscores the necessity for a complete structured framework that combines modern Natural Language Processing Approach with Tf-idf method, hybrid model and visualization tools to properly evaluate the comments from YouTube e-videos. This research study proposes a solution to all these challenges, offering an innovative approach to improve educational content on YouTube using sentiment analysis.

Below table brief the details of some of the reviewed papers:

Table 1: Summary Table of Reviewed papers

Year	Author(s)	Domain	Methodology Applied	Challenges	Accuracy
2022	Rajesh and Akila	E-learning videos	Naïve Bayes and SVM Hybrid model	Struggled with nuanced sentiments like sarcasm and mixed emotions	89%
2021	Singh and Tiwari	YouTube comments	Tokenization	Difficulty in handling polarity and slang	82%
2017	Shiha and Avyaz	Emojis in SA	Emoji sentiment mapping	Focused on social media; did not address educational video content	85%
2024	Maw et al.	YouTube comments	Deep learning (BiLSTM)	Computational inefficiency and interpretability issues	92%
2019	Xu et al.	Chinese Text SA	BiLSTM, Domain-specific Dictionaries	Limited to standard text, not multilingual or informal	88%
2020	Rajeswari et al.	Customer Sentiments	Hybrid Models	Educational feedback was not addressed	87%
2019	Wang et al.	Twitter sentiment	SentiDiff Model	Limited scalability to educational platforms	90%
2021	Naresh Kumar and Uma	Social Media sentiments	Lexicon Based and Context Aware Model	Challenges with domain specific lexicons for educational context	86%

3 Research Methodology

This research work employs Data mining and Natural language processing to analyse sentiments of the user comments on YouTube educational content. The aim of the project is to identify gaps and key areas to improve and innovate the research.
We will be using KDD approach to structure the methodology of our research as shown in Fig.1. It will explain the architectural, scientific methodology and design from beginning to the completion of the research.

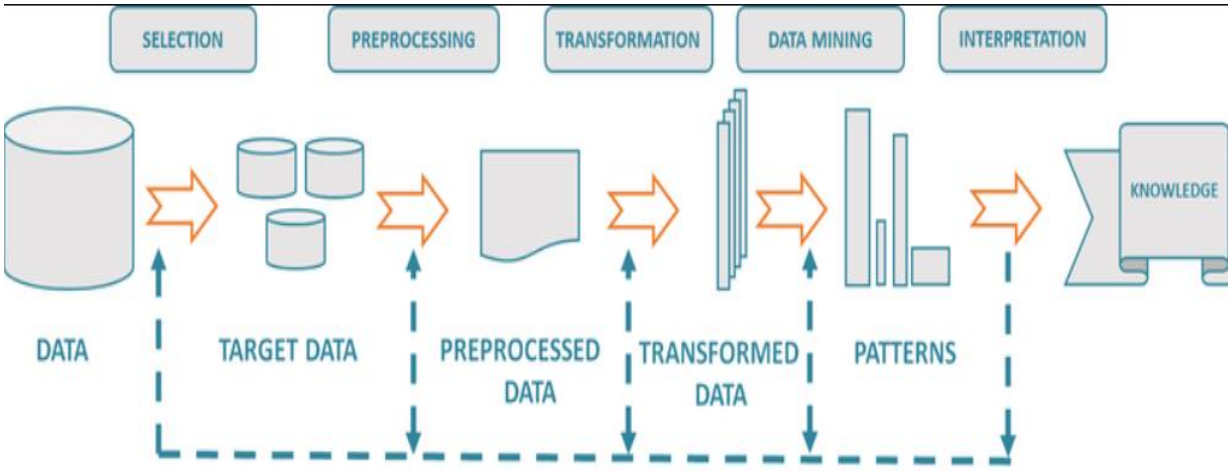

Fig.1. Step-by-Step KDD Methodology

This section of research will emphasize the steps and approaches involved in achieving the desired results.

Data Source:
We have extracted our data from Youtube video with the tagline as "How I'd Learn AI in 2024 (if I could start over)" by "Dave Ebbelaar"
URL: "https://www.youtube.com/watch?v=h2FDq3agImI&t=85s". There are altogether 643 comments and 5 columns in the dataset.

The following steps elaborate about the process of research findings:

1.Data Collection
This data is identified and collected after many relevant analyses of the video specifically an educational video and sentiment analysis is applied on it.
YouTube Data API v3 is the data source platform which allowed to collect the public comments which are freely available.

Method: We used numerous python libraries and mainly the one which interact with the API and extract comments from the video i.e. **googleapiclient.discovery**. This library extract data including metadata such as author, date, published and likes.
While, collecting the data we ensured that the comments are up-to-date, representative of the video's audience and relevant.

2. Data Preprocessing

The data preprocessing step prepares the raw data for further analysis by addressing the noise and inconsistencies in the data. It is divided into two parts that is Data Cleaning and Feature Transformation.

Data Cleaning:
In this process we focus on standardizing the data by implementing text normalization techniques.
. **Lowercasing**: Converted all the text to lowercase to make sure the text in uniform all over.

. Punctuation Removal: Elimination of punctuation marks and special characters using regular expressions.

	like_count	text
0	132	thank half million view bryou find free roadma...
1	0	necessary learn dsa c well
2	0	beautiful
3	0	good video people already started watched vide...
4	0	

Fig.2. Preprocessed comments

. Tokenization: Divided the text into discrete words

. Lemmatization: Reduced words to their base form by using nltk and TextBlob.

. Stopword Removal: Removed common English stop words to focus on meaningful words

Data Transformation:

By using TfidfVectorizer to build a feature matrix that represents textual data. Converted the textual data into numerical format to understand common two-word phrases and individual words by capturing both bigrams and unigrams.

Verified at end of all this process that the contextual meaning of the comments is not lost and ensure the preprocessed data maintains good representation of the original extracted comments.

3.Data Mining

As part of this process, applied machine learning algorithms to classify the sentiments in the data fetched and extract patterns.

. Sentiment Analysis:
To classify the sentiment is positive, negative or neutral and to calculate polarity; TextBlob is used.
Applied custom logic to derive sentiment categories to map the polarity scores on the basis of predefined thresholds

i.e. >0.1 for positive, -0.1 to 0.1 for neutral and <-0.1 for negative).

. Model Selection and Training:
Random Forest Classifier model is selected due to its ability of robust performance and handling high dimensional data for text classification step.
Splitted the dataset into training and testing subsets to evaluate effectiveness of the model.
Fitted the model on training data and validated using cross-validation.

. Evaluation Metrics:
Performance of the trained model is measured using metrics such as accuracy, precision, recall and F1Score with the results.

4. Data Evaluation and Interpretation

Interpreted the data mining results by the model's performance and sentiment distribution factor.
Confusion Matrix with all metrics like precision, recall and F1-Score for sentiments is been analysed and reviewed.
Visualization with the help of polarity graph of the sentiment distribution from positive to negative and box-plot categorised into the scale of positive, negative and neutral comments for the video.
Generated a word cloud to visually represent frequently occurring words in comments.

4 Design Specification

The comprehensive understanding of the research work is designed in this section which includes several key components and steps. It defines the architectural flow of the work.

1. Framework and Architecture

. Data Pipeline and Preprocessing Module:

To collect comments from YouTube the system uses YouTube API and then those extracted comments are processed using different Python libraries useful in the course of work. Libraries like pandas is utilized for data manipulation and **nltk** for processing of text by all sub-processes like lemmatization, stop word removal. Feature Extraction is done by using **Tf-idf Vectorizer.**

. Sentiment Analysis Module:

A sentiment classifier i.e. Random Forest Classifier is used to train on the pre-processed data with labels taken from **TextBlob** polarity scores.
The results of this are formulated by using evaluation metrics by integrating accuracy, precision, recall, F1-score and visualization of the confusion matrix.

. Visualization and Interpretation Module:

Use of other libraries like **seaborn**, **matplotlib**, and **wordcloud** is done to generate word clouds and plots. Polarity graphs are created to display trends across the comments.

2. Requirements

. Software:
Access to YouTube API for collection of data and also different python libraries are used like **googleapiclient, textblob, pandas, nltk, sklearn, seaborn, matplotlib** and **wordcloud**.

. Hardware:
A system is efficient to handle TF-IDF computation and Random Forest training as it is with 8GB RAM and multi-core processor.

. Data:
Access to all comments of diverse educational videos; ensuring that the results are generalized.

Functionality and workflow:
1.Comments are extracted from YouTube videos using the YouTube API.
Then they are structured into a dataset frame and saved in a CSV file at desired location.

2.Preprocessing of the extracted data is then carried out which ensures the data text cleanliness and transformed comments into usable format by using TF-IDF.

3.The polarity of the sentiments is classified by use of TextBlob and also a refined results for labelled data is classified with Random Forest Classifier.

4.To gain insights into sentiment trends and frequencies of word; Visualizations are generated in form of polarity graphs, bar graphs and word cloud.
All this process ensures modularity and adaptability. Below figure 3 demonstrates the workflow in brief.

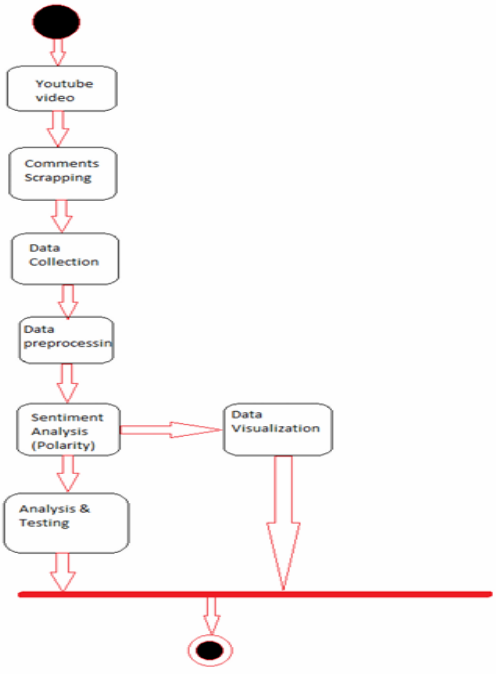

Fig.3. Workflow Flowchart

5 Implementation

This section elaborates the use of all techniques and transformation for the research work. Process uses Natural Language processing (NLP) techniques, machine learning model and interpretation and visualization tool.

. **Natural Language Processing (NLP)** is preprocessing technique that pre-process the human language data into computer understandable language for analysis. The human language data which is textual data is always a raw data which needs to be cleaned up so that it is ready for further process. NLTK which is the python package also known as toolkit of NLP. To use this NLTK library we have to install it in python environment and it must be imported into Jupyter notebook.
. **Sentiment Analysis** is a technique part of Natural language processing which is used to label data in form of classified categories. The sentiment analysis process helps understanding the emotions behind the textual data we are analysing by categorizing it on basis of parameters.
Focusing on the technologies used and output produced we have detailed explanation of the implementation stage of the research as below.

Overview:
The aim of the implementation is to analyse the comments extracted from the YouTube educational video and apply sentiment analysis to the fetched comments.
The process starts from fetching the comments from the specific desired video using the API i.e. YouTube API. The fetched comments are then pre-processed using multiple techniques for sending them to next step that is sentiment analysis.TF-IDF vectorization is applied to extract features. After feature extraction, a machine learning model is introduced for sentiment classification which is then trained and evaluated. The output of the process till now is then visualized through word clouds and graphs.
The output of this process includes below points:
. Cleaned, transformed and pre-processed comments dataset.
. Comments being represented as feature vectors.
. Sentiment polarity scores in tabular format with classified categories like positive, neutral and negative.
. Evaluation metrics of the machine learning model used.
. Visualizations in form of polarity trends, sentiment distributions and commonly used words in form of word cloud.

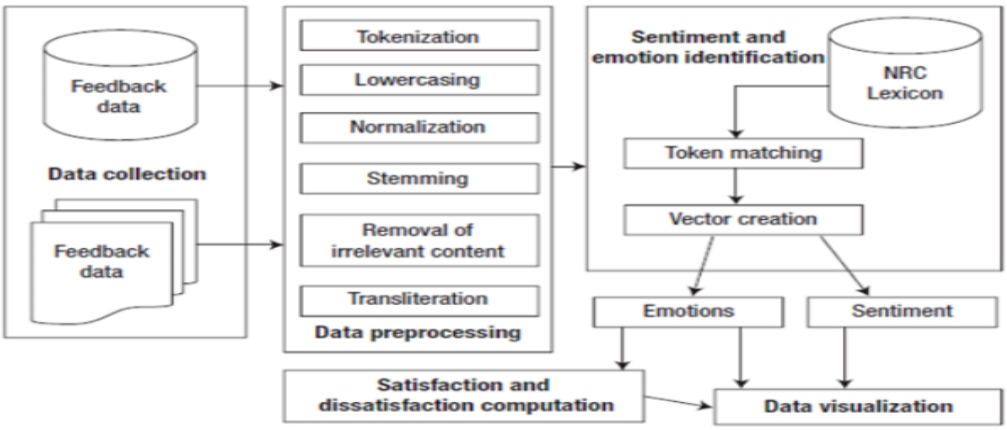

Fig.4. System Architecture

Tools and Libraries Utilized for the research:

1. Data Extraction and Manipulation:
. Google API Client i.e. YouTube Data API to extract and fetch the comments from the desired video.
. Pandas, which is used for data manipulation and cleaning in tabular form.

2. Data Pre-Processing:
. NLTK used for preprocessing steps which includes stop word removal, lemmatization and tokenization.
. TextBlob used for analysis of the sentiments of the comments retrieved and also for calculating the polarity of the fetched sentiments.

3. Machine Learning Model:
. For splitting the dataset, training models, transformation of the text into feature vectors and evaluating performance; Scikit-learn is used.
. To convert text data into numerical features we use TF-IDF Vectorizer
which is the term frequency inverse document frequency method.
. The chosen supervised learning model Random Forest Classifier is then used for prediction of the sentiments.

4. Interpretation and Visualization:
Evaluation metrics for interpretating the results in format and Matplotlib for creating polarity graphs and seaborn for sentiment distribution across and Word Cloud for generating graphical representation of frequently occurring words in the comment section.

. Implementation in Detail

Fetching and Saving YouTube Comments from Educational video:
The process starts from retrieving comments from specified video and they are fetched in batches with the use of YouTube Data API. These comments include all attributes like name of the author, publication date, number of likes on individual comment and the comment text. These are then stored in a structured tabular format using Pandas libraries and saved as a CSV file for further analysis.
Output: A CSV file is saved which contains all raw comments along with metadata.

Preprocessing of Text Data:
This step involves cleaning and transforming the text for further subsequent analysis.
It applies functions like lowercasing which standardize the text to lowercase for uniformity.
Punctuation Removal which means eliminating non-alphanumeric characters to simplify tokens for further process.
Removing stop words i.e. "the", "is", and "and" using NLTK library.
Reducing the words to their base form using lemmatization.
The pre-processed comments are then stored in Pandas Data frame for subsequent analysis.
Output: A pre-processed data column in the dataset is obtained which is ready for feature extraction.

The pre-processed data after applying all the techniques is demonstrated in below figure 5.

```
Preprocessed Data:
                author           published_at           updated_at  \
0         @daveebbelaar  2023-08-04T17:23:55Z  2024-03-27T12:12:17Z
1  @RadhaKrishnaRangoli  2024-12-04T09:18:56Z  2024-12-04T09:18:56Z
2            @Mahdi-s3i  2024-12-04T08:10:14Z  2024-12-04T08:10:14Z
3     @chillmusiclyrics  2024-12-03T11:12:36Z  2024-12-03T11:12:36Z
4         @JeseleeLewis  2024-12-01T19:32:58Z  2024-12-01T19:32:58Z

   like_count                                               text  \
0         132  thank half million view bryou find free roadma...
1           0                          necessary learn dsa c well
2           0                                           beautiful
3           0  good video people already started watched vide...
4           0

   sentiment_polarity predicted_sentiment actual_sentiment
0            0.116667            positive         positive
1            0.000000             neutral         negative
2            0.850000            positive          neutral
3            0.134091            positive         positive
4            0.000000             neutral         negative
```

Fig.5. Preprocessed Comments Data

Feature Extraction Using TF-IDF Vector:

This step involves converting the pre-processed comments from above step into numerical feature vectors by use of TF-IDF vectorization. These vectors work by capturing the words that are important from the dataset by considering their frequency across entire dataset frame and also individual comments.

. Both bigrams and unigrams are used to capture the sequence of the words.

. Dimensionality Reduction is carried out to reduce computational complexity by limiting the vocabulary to upto top 1000 features.

Output: A sparse matrix is obtained which represents comments as TF-IDF feature vectors.

Classification of Sentiments:

This step classifies the obtained comments using Text Blob by assigning a polarity score to each textual comment. It is then categorized as:

. Positive: Polarity > 0.1

. Neutral: Polarity between -0.1 and 0.1

. Negative: Polarity < -0.1

After this step, the dataset is then labelled with a hypothetical sentiments known as "actual sentiments" to enable supervised learning. These labels help simulating real-world data.

Output: Polarity Score is derived for each comment from the data and also sentiment labels are predicted i.e. positive, negative or neutral.

Machine Learning Model:

•Random Forest Classifier

Random Forest is one of the versatile and easy-to-implement algorithm.

In this algorithm multiple decision trees are combined together to create a random forest. The accuracy increases if the number of trees in the random forest classifier increases.

Use of Random Forest Classifier is done which is trained on the labelled dataset obtained from above steps using TF-IDF features. This model splits the data into training 80% and testing 20% sets for further evaluation process. Evaluation metrics like accuracy, precision, F1-score are calculated to know the performance of the model.

Output: The used machine learning model's performance metrics is derived and test data is predicted.

Visualization:
. A Sentiment Polarity Graph is plotted which is a line plot and displays the polarity scores of fetched comments where red line indicates that the comment is neutral.
. A Sentiment Distribution bar graph which shows the frequency of positive, negative and neutral sentiment comments from the data.
. A graphical representation of recurring used words from the dataset is shown in form of a Word Cloud.
Output: Visualized the polarity trends in comments, distribution of the categorical comments sentiment and also gain insights into common terms used in the comments from the audience.

Key Outcomes and Results:
. The Random Forest model achieved average accuracy, precision, recall and F1-score as seen in the classification report.
. Positive comments are more prevalent than negative comments.
. Neutral comments are too in significant amount.
. Highlighted key topics discussed in the comment section through word cloud analysis.

6 Evaluation

In this section, the evaluation of the machine learning model is done out with classifying the evaluation metrics parameters like accuracy, precision score, recall, F1-Score.
The results do show room for improvements. Following are the metrics as obtained:

1.Accuracy and Overall Performance:
. The model achieved an accuracy of 33% which is low for classification of three-class categorical task of positive, negative and neutral.
This accuracy value suggest that the predictions of the model are better than guessing randomly.

2. Precision, F1-Score and Recall, Macro and Weighted Averages:
. Positive Class-
Precision: 0.34 - indicates that out of all our predictions as positive only 34% were correct.
F1-Score: 0.35 - Balances precision and recall by highlighting the inconsistencies in prediction class.
Recall: 0.37 - Suggest that of positive samples 37% were correctly identified.

. Negative Class -
Precision: 0.41, Recall: 0.37, F1-score: 0.39
This is similar to positive class predictions but the positive sentiments are marginally better than negative but overall still unsatisfactory.

. Neutral Class-
Precision: 0.24, Recall: 0.24, F1-score: 0.24 are the neutral class predictions which indicates low performance detecting neutral sentiments.

. Macro and weighted Averages -
Macro averages all over is 0.33 that the model struggles equally around all the classes
Weighted averages show no class as favourite as all has same performance. As shown in figure 6.

```
Accuracy: 0.33
Precision: 0.34
Recall: 0.33
F1 Score: 0.33
```

```
# Detailed Classification Report
print("\nClassification Report:\n")
print(classification_report(y_test, y_pred))
```

```
Classification Report:

              precision    recall  f1-score   support

    negative       0.41      0.37      0.39        30
     neutral       0.24      0.24      0.24        25
    positive       0.34      0.37      0.35        38

    accuracy                           0.33        93
   macro avg       0.33      0.33      0.33        93
weighted avg       0.34      0.33      0.33        93
```

Fig.6. Evaluation metrics and classification report

Visualization:

1. Sentiment Distribution Bar Graph

. The X-axis has sentiment categories like 'Positive', 'Negative' and 'Neutral'. across all the comments fetched.

. Y-axis counts the number of comments falling in each category.

We could see that positive comments are more frequent, suggesting the reception of video by the students or users in favourable way.

Neutral comments also have significant part which indicates the tone of comments unopinionated.

Negative comments are rare, suggesting minimum dissatisfaction among the video viewers.

The below graph in Figure 7 demonstrates it in detail.

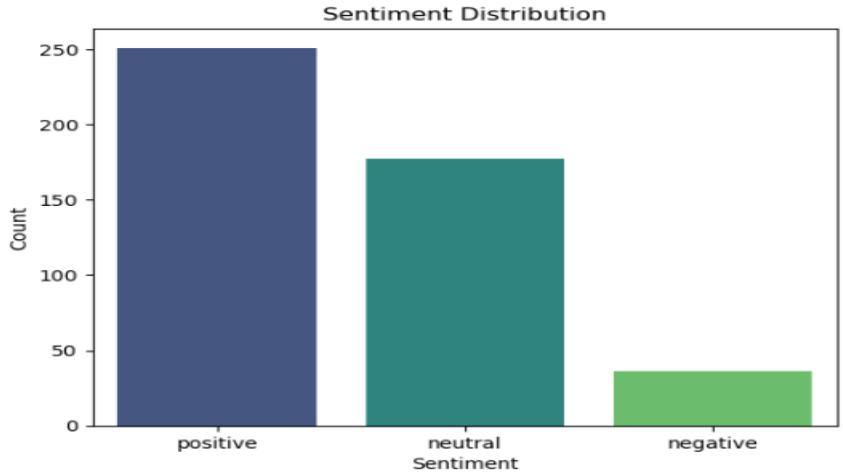

Fig.7. Bar Graph of Sentiment Distribution

2. Sentiment Polarity Graph

. The X-axis indicates comment index which represents the sequential order of comments from the dataset.

. The Y-axis defines the polarity score which is sentiment score ranging from -1 i.e. strongly negative to +1 i.e. strongly positive.

The red dashed line indicates neutral polarity and the blue line indicates polarity values of individual fetched comments from the video. As indicated in below polarity graph in Fig. 8.

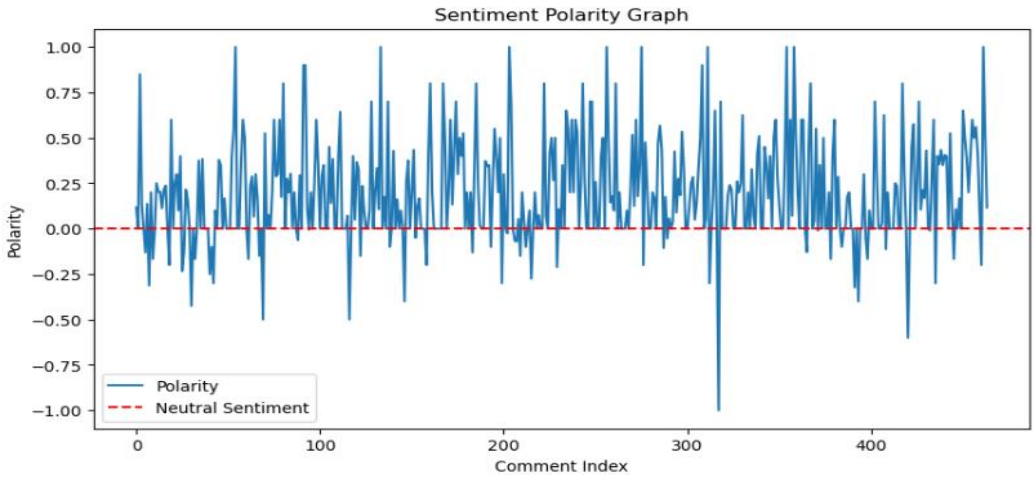

Fig.8. Sentiment Polarity Graph

There is variation seen in sentiment polarity as it ranges widely from positive to negative and have spikes accordingly.

The majority of comments lie near the neutral range and some show peak strongly negative or positive sentiments.

3. Sentiment WordCloud

. The words like "thank" and "video" appear more prominently, which shows people show gratitude towards the video.

. The words like "ai", "learn", "python", "learning", reflects the educational focus on artificial intelligence and programming with python.

. The words like "need" and "want" reflects viewers expressing some requirements.

Fig.9. Word Cloud of frequent words

Discussion

The results from the evaluation define that the current approach and design of the sentiment analysis process has several limitations.
. The Random Forest model which is robust for structured data is not good for classification of text. Models like SVMs, logistic regression, deep learning models like BERT can perform better due to their ability to handle complexity in textual data relationships.
. The TF-IDF vectorization measures frequency count of words but fails to look at semantic and contextual nuances, which are important for sentiment analysis.

To improve the design of the research, incorporating contextual embeddings like BERT or GPT-derived embeddings can be used and also improvements in handling sarcasm and negations while preprocessing could produce better results. These findings emphasize on the necessity for contextualized text representation for sentiment analysis.

7 Conclusion and Future Work

The purpose of this research work is to classify YouTube comments into positive, negative and neutral comments based on the sentiments expressed in the text using TF-IDF based feature extraction and Random Forest classifier data mining method. While the implement process is successful in processing the data, making predictions and finding results with what it was desired at the start of the process. But it has revealed significant limitations. Where the model attained 33% accuracy with low precision, recall and F1-Scores values consistently across all sentiment classes. These overall findings emphasize on the point that traditional machine learning approaches have limitations in handling text input data that is nuanced and contextual in nature.

This research emphasizes the necessity for more advanced feature extraction methods and model architectures in order to achieve reliable sentiment analysis outcomes.

To identify and work on these shortcomings, future work should use BeRT or GPT which make use of contextualized embeddings to capture semantic relationships in text data. Also, addressing the imbalance of the dataset by class weighting or oversampling approach. As well as applying advanced preprocessing approaches to detect sarcasm; might improve the performance of the model. Hybrid approaches like deep learning and lexicon-based methods can also be explored to know their outcomes and findings.

These improvements could result in more accurate and commercially feasible sentiment analysis system

8 Acknowledgement

My sincere thanks to the project supervisor Prof. Vladimir Milosavljevic for giving me the knowledge and independence to work on my thesis through his guidance over the past weeks to complete this thesis successfully. I would also like to express my gratitude towards MSc. Data Analytics department to provide me with report writing related information and the National College of Ireland for allowing me to show my skills throughout.

References

P. Rajesh and D. Akila, 2022, "Sentimental analysis on E-Learning videos using Hybrid Algorithm based on Naïve Bayes and SVM," 2022 International Conference on Emerging Smart Computing and Informatics (ESCI), Pune, India, pp. 1-6.

Singh, R. and Tiwari, A., 2021. Youtube comments sentiment analysis. International Journal of Scientific Research in Engineering and Management (IJSREM), 5(5), pp.1-11.

Shiha, M. and Ayvaz, S., 2017. The effects of emoji in sentiment analysis. Int. J. Comput. Electr. Eng.(IJCEE.), 9(1), pp.360-369.

B S. S. Maw, E. C. Lwin, W. Mar, N. S. Paw, M. M. Khaing and T. T. Aung, 2024,"Sentiment Analysis with YouTube Comments Using Deep Learning Approaches," 2024 IEEE Conference on Computer Applications (ICCA), Yangon, Myanmar, pp. 1-7.

Xu, G., Yu, Z., Yao, H., Li, F., Meng, Y. and Wu, X., 2019. Chinese text sentiment analysis based on extended sentiment dictionary. IEEE access, 7, pp.43749-43762.

Pooja and Bhalla, R., 2022. A review paper on the role of sentiment analysis in quality education. SN Computer Science, 3(6), p.469.

Wang, L., Niu, J. and Yu, S., 2019. SentiDiff: combining textual information and sentiment diffusion patterns for Twitter sentiment analysis. IEEE Transactions on Knowledge and Data Engineering, 32(10), pp.2026-2039.

Shoufan, A. and Mohamed, F., 2022. YouTube and education: A scoping review. IEEE Access, 10, pp.125576-125599.

Pai, A.R., Prince, M. and Prasannakumar, C.V., 2022, June. Real-time twitter sentiment analytics and visualization using vader. In 2022 2nd International Conference on Intelligent Technologies (CONIT), pp. 1-4.

Naresh Kumar, K.E. and Uma, V., 2021. Intelligent sentinet-based lexicon for context-aware sentiment analysis: optimized neural network for sentiment classification on social media. The Journal of Supercomputing, 77(11), pp.12801-12825.

Saif Mohammad and Peter D Turney, “NRC emotion lexicon”, National Research Council Canada (2013), pp 234.

Athindran, N.S., Manikandaraj, S. and Kamaleshwar, R., 2018, November. Comparative analysis of customer sentiments on competing brands using hybrid model approach. In 2018 3rd International Conference on Inventive Computation Technologies (ICICT), pp. 348-353.

Cheng, L.C. and Tsai, S.L., 2019, August. Deep learning for automated sentiment analysis of social media. In Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining, pp. 1001-1004.

Jain, G., Verma, S., Gupta, H., Jindal, S., Rawat, M. and Kumar, K., 2022, July. Machine Learning Algorithm Based Emotion Detection System. In 2022 Fifth International Conference on Computational Intelligence and Communication Technologies (CCICT), pp. 270-274.

Rajeswari, A.M., Mahalakshmi, M., Nithyashree, R. and Nalini, G., 2020, July. Sentiment analysis for predicting customer reviews using a hybrid approach. In 2020 Advanced Computing and Communication Technologies for High Performance Applications (ACCTHPA), pp. 200-205.

Saberi, B. and Saad, S., 2017. Sentiment analysis or opinion mining: A review. Int. J. Adv. Sci. Eng. Inf. Technol, 7(5), pp.1660-1666.

Mujahid, M., Lee, E., Rustam, F., Washington, P.B., Ullah, S., Reshi, A.A. and Ashraf, I., 2021. Sentiment analysis and topic modeling on tweets about online education during COVID-19. Applied Sciences, 11(18), p.8438.

Roy, A. and Ojha, M., 2020, December. Twitter sentiment analysis using deep learning models. In 2020 IEEE 17th India council international conference (INDICON), pp. 1-6.

Wang, M. and Yu, Y., 2022, August. Deep sentiment analysis of the feelings expressed by tourists based on bert model. In 2022 International Conference on Culture-Oriented Science and Technology (CoST), pp. 130-133.

Xu, G., Meng, Y., Qiu, X., Yu, Z. and Wu, X., 2019. Sentiment analysis of comment texts based on BiLSTM. Ieee Access, 7, pp.51522-51532.