BMS INSTITUTE OF TECHNOLOGY AND MANAGEMENT

YELAHANKA, BENGALURU - 560064

DEPARTMENT OF INFORMATION SCIENCE AND ENGINEERING

PROJECT BASED LEARNING

Odd Semester 2023-24

Synopsis of

"Sentiment Analysis"

Data Structures – BCS304

III Semester

Section - B

Submitted By

Raihan Batool	USN: 1BY22IS127
Netravavti Ramachandra Hegde	USN: 1BY22IS100
Lakshmi Priya Padhihari	USN: 1BY22IS072
Meghana Purumu	USN: 1BY22IS087

Under the Guidance of

Dr. Karthik S A
Assistant Professor

Dr. Anil Kumar
Assistant Professor

2023-2024

TABLE OF CONTENTS

ABSTRACT

Sentiment analysis is a technique that uses natural language processing (NLP) to determine emotions or sentiments expressed in a piece of text. User reviews or Viewer comments are one of the best ways to understand what's happening or what the public thinks of certain things. YouTube is one of the most popular videos sharing platforms obtaining millions of views and receive several comments. These comments often contain valuable information that help in improving the rating levels of the uploaded content. The comments can be analyzed by using NLP techniques and machine learning techniques and be classified as negative, neutral or positive.

In our project, we aim to accurately predict viewer's sentiment by analyzing the data provided using the Naïve Bayes Classifier Algorithm. The project uses the Google Developer's YouTube Application Program Interface (API) key to extract comments. The comments from a YouTube video are first extracted using the Google Developer's YouTube API key. The comments are then cleaned to remove any unwanted text such as emoticons before being analyzed. The project demonstrates the potential of sentiment analysis for understanding the opinions and preferences of the YouTube audience.

INTRODUCTION

In the digital age, where opinions echo through the vast expanse of the internet, user reviews and public comments have emerged as invaluable glimpses into the collective consciousness. The virtual realm, especially platforms like YouTube and Twitter, serves as a digital agora where users express thoughts, critiques, and sentiments about an array of topics ranging from entertainment to current affairs. Understanding the pulse of this digital chatter has become crucial for deciphering public sentiment.

Sentiment analysis, also known as opinion mining, is a field of natural language processing (NLP) that aims to automatically identify and extract subjective information from texts. Sentiment analysis has numerous applications, and can be applied to various domains and tasks, ranging from marketing to politics. It has become an increasingly popular topic of research in the past decade. In this project, we are going to perform sentiment analysis on YouTube comments that we have collected using the YouTube API. Despite recent advancements, the detection and analysis of sentiments remain challenging due to several factors. One of the most significant challenges in sentiment analysis is the ambiguity of language. Another challenge in sentiment analysis is the detection of tone and sarcasm. Texts often contain tones that can be difficult for algorithms to detect, and sarcasm can be especially challenging. Cultural and contextual differences make things more complex. Feelings can change based on culture and context, so what seems positive in one culture might not have the same meaning in another. This project sets the scene for exploring sentiment analysis, looking at its complexities, advancements, and ongoing challenges.

MOTIVATION

User-generated content has evolved into a powerful force that shapes perceptions, influences decisions, and mirrors the collective consciousness of society. Among the myriad forms of user engagement, YouTube comments stand out as a rich source of unfiltered, spontaneous reactions to a diverse range of content. These comments are more than just text; they represent a treasure trove of authentic, raw expressions of sentiment. Hence, sentiment analysis is useful for quickly unravelling the collective psyche of the digital society. Predicting sentiments – positive, negative, or neutral – within this vast pool of interactions unveils a nuanced understanding of human expression. This synthesis of technology and emotion becomes a valuable tool for providing a preliminary examination of products, movies, and various subjects that captivate the online community.

In the realm of sentiment analysis, YouTube comments offer a unique terrain for exploring the intersection of machine learning, data science, and linguistics. As we endeavour to train models to comprehend the emotional nuances within these comments, we push the boundaries of what is achievable in the field of machine learning. The technical motivation is not merely about decoding sentiments but also about creating systems that continuously learn and adapt to the evolving nature of online communication. Crafting algorithms that can navigate this linguistic diversity is not just a pursuit of technological advancement but a gateway to understanding the intricacies of human communication on a global scale.

EXISTING SYSTEM

The field of sentiment analysis has witnessed significant advancements, and several approaches are currently employed to analyze and understand sentiments in textual data. Some of the existing systems for sentiment analysis are:

Twitter-roBERTa-base:

Twitter-roBERTa-base was developed by researchers from Cardiff University and is based on the original RoBERTa-base model by Facebook. It is a language model that was trained on about 58 million tweets and fine-tuned for various natural language processing tasks, such as sentiment analysis, emotion recognition and irony detection.

EffectCheck:

EffectCheck is a software tool developed by Effct Technologies, Inc., that analyzes the emotional impact of words and phrases in a piece of text. The software allows the user to calculate the emotional score on a piece of text, tailor the message to fit the specific audience, and helps optimize the impact of that writing.

LIMITATIONS OF EXISTING SYSTEM

Despite recent advancements, the domain of sentiment analysis remains challenging due to several factors. Besides, it is challenging to choose one best accurate model.

1. **Vagueness and Ambiguity in Sentences:** The existing system grapples with sentences or words that are often vague and difficult to identify, introducing a challenge in accurately analysing targeted data. The ambiguity in expressions may lead to unreliable outcomes in sentiment analysis.
2. **Difficulty in Understanding Opinions:** The system encounters difficulty in precisely understanding the intended meaning of opinions expressed in some sentences. For instance, when faced with sentences like "glad to expose the disciplinary committee steroids of player Noor," the exact target of the sentiment may remain unclear. The system may struggle to distinguish between positive and negative sentiments, especially when the same event is viewed differently by individuals supporting different sides.
3. **Neutral Reports of Events:** Neutral reports of events pose a challenge, as the emotional state of the speaker is not always clear before describing situations. Deciphering whether reports are emotionally charged or presuppose a negative emotional state can be challenging for sentiment analysis.
4. **Detection of Abusive Opinions and Counterfeit Reviews:** The system faces challenges in identifying and filtering out spam information, including abusive opinions and counterfeit reviews. The presence of such unwanted messages can impact the reliability of sentiment analysis results.
5. **Domain-Specific Focus:** The system tends to focus on one domain, limiting its adaptability to diverse topics. The nature of sentiment analysis may lead to a concentration on a specific subject, potentially yielding biased results that may not be applicable across different domains.
6. **Expensive Opinion Mining Software:** The high cost of opinion mining software poses a significant limitation. The existing system struggles with the financial barrier, as these tools are expensive and often only accessible to governments and large organizations. This limitation hinders widespread adoption, preventing the average person from benefiting from sentiment analysis software.

PROPOSED SYSTEM

The proposed system is a web-based application that performs sentiment analysis on YouTube comments. The system uses takes YouTube comments as input, cleans them by removing noise and irrelevant words, and analyzes them by using a machine learning algorithm to classify them into positive, negative, or neutral categories. The system also displays the percentage of each category, as well as the overall sentiment score of the video. The system also identifies the most positive and most negative comments for further analysis.

The system tackles the following issues:

1. **Scalability and Performance Optimization:** The system is designed to scale seamlessly with growing traffic and data.
2. **Responsive and Intuitive User Interface:** The system has a user-friendly interface and prioritizes intuitive navigation and a visually appealing layout.
3. **Nuances of Comments:** It improves the accuracy and reliability of sentiment analysis by using a machine learning algorithm that can capture the linguistic features and nuances of YouTube comments, such as slang, abbreviations, emojis, and hashtags.

SYSTEM REQUIREMENT SPECIFICATIONS

Tools and Frameworks

Python 3.11- boasts a rich ecosystem of libraries and frameworks tailored for machine learning, scientific computing, data manipulation and visualization

Google developer's YouTube data API - a set of web services that allow you to access and manipulate YouTube data, such as videos, channels, playlists, comments, and more

Naïve Bayes Classifier - a simple and fast machine learning algorithm to predict the class of a new data point based on calculated probabilities

Binary Search Tree - a tree data structure where each node can have at most two children, a binary search tree is used to take comments from a file and clean the comment by removing unwanted words

urllib – for working with URLs

apiclient - for creating and managing API client libraries with urllib

csv - for reading and writing data in CSV format

json - for encoding and decoding data in JSON format

matplotlib – for creating output visualizations

Functional requirements

Functional requirements specify the features, functionalities, and behaviours that a system must have to satisfy user needs. They focus on what the system should do and how it should respond to user actions.

1. Interactive environment
2. Output visualization

PROPOSED METHODOLOGY

The proposed methodology for this project consists of the following steps:

Data collection

YouTube comments are collected from various videos using the Google Developer's YouTube API. The API calls are made using the API key that we have generated and the 'urllib' library in Python. Then, we specify the video ID, the number of comments, and the fields that we want to retrieve, such as the comment text, the author's name, the like count, and the reply count.

Data preprocessing

The comments then undergo preprocessing by applying various techniques to improve the quality and readability of the data. The NLTK library in Python is used to perform the following tasks:

1. **Tokenization:** The comments are split into individual words or tokens, and any punctuation marks or symbols that are not part of the words are removed.
2. **Lemmatization:** The split words are converted into their base or dictionary form, by using the WordNetLemmatizer class in NLTK. This will help us reduce the dimensionality and variability of the data, and improve the accuracy of the analysis.
3. **Stopword removal:** Very common words that do not carry much meaning or sentiment, such as "the", "a", "and", etc. are removed. This is done using a predefined list of stopwords in NLTK. Along with these words, some custom stopwords that are specific to YouTube comments, such as "lol", "omg", "wtf", etc. are also removed.
4. **Noise removal:** Any words or characters that are irrelevant or redundant, such as URLs, emojis, hashtags, mentions, etc. are identified and removed using regular expressions and string methods in Python.

Data analysis

Then we perform sentiment analysis on the preprocessed comments by using a machine learning algorithm. Optionally, we can use the TensorFlow library in Python to build and train a deep neural network that can classify the comments into positive, negative, or neutral categories.

Data visualization

The results of the sentiment analysis are visualized by using various charts and graphs, such as pie charts, bar charts, histograms, etc. Creation and display of the visualizations is done using the matplotlib and seaborn libraries in Python to create and display the visualizations.

REFERNCES

1. R. F. Alhujaili and W. M. S. Yafooz, "Sentiment Analysis for Youtube Videos with User Comments: Review," 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), Coimbatore, India, 2021, pp. 814-820, doi: 10.1109/ICAIS50930.2021.9396049.
2. A. Gkillas, M. A. Simos and C. Makris, "Popularity Inference Based on Semantic Sentiment Analysis of YouTube Video Comments," 2023 14th International Conference on Information, Intelligence, Systems & Applications (IISA), Volos, Greece, 2023, pp. 1-6, doi: 10.1109/IISA59645.2023.10345963.
3. Sindhu, S. Kumar and A. Noliya, "A Review on Sentiment Analysis using Machine Learning," 2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA), Uttarakhand, India, 2023, pp. 138-142, doi: 10.1109/ICIDCA56705.2023.10099665.
4. H. Chauhan, K. S. Rathore and V. G. Rajan, "Sentiment Analysis using Supervised Machine Learning," 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2022, pp. 1276-1279, doi: 10.1109/ICACITE53722.2022.9823681.
5. B. D. Shivahare, S. Ranjan, A. M. Rao, J. Balaji, D. Dattattrey and M. Arham, "Survey Paper: Study of Sentiment Analysis and Machine Translation using Natural Language Processing and its Applications," 2022 3rd International Conference on Intelligent Engineering and Management (ICIEM), London, United Kingdom, 2022, pp. 652-656, doi: 10.1109/ICIEM54221.2022.9853044.
6. Y. Chandra and A. Jana, "Sentiment Analysis using Machine Learning and Deep Learning," 2020 7th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 2020, pp. 1-4, doi: 10.23919/INDIACom49435.2020.9083703.
7. Abdellatif, Abdelrahman, Sahmoud, Shaaban & Nizam, Ali., “A Unified Framework for Multi-Language Sentiment Analysis,” 2023 3rd International Conference on Computing and Information Technology (ICCIT), Tabuk, Kingdom of Saudi Arabia, 2023, pp. 280-284, doi: 10.1109/ICCIT58132.2023.10273894.