CSCE 633:FALL 2023 Final Report - Ankit Modi

YouTube Video Title Sentiment Analysis

Goals

This report talks about the importance of sentiment analysis on YouTube video titles, how it can be useful, the analysis on YouTube Trending Video dataset, and the BERT (Bidirectional Encoder Representations from Transformers) model that is used for sentiment analysis.

Introduction

In recent years, sentiment analysis has gained significant attention due to its applications in understanding public opinion, customer feedback, and social media content. Sentiment analysis involves determining the sentiment or emotional tone expressed in a piece of text. With the advent of powerful language models like BERT (Bidirectional Encoder Representations from Transformers), the field has seen substantial improvements in accuracy and efficiency.

BERT, introduced by Google, is a transformer-based model that has achieved state-of-the-art results in various natural language processing (NLP) tasks, including sentiment analysis. Unlike traditional models, BERT considers the context of words bidirectionally, capturing dependencies and relationships more effectively.

Sentiment analysis on YouTube video titles plays a pivotal role in determining the appropriateness and accessibility of content for a diverse audience. The sentiment expressed in video titles serves as a crucial indicator of the emotional tone and potential impact on viewers. Here's how sentiment categorization influences the accessibility of YouTube content:

- Positive Sentiment:
  - For content creators, incorporating positive sentiments in video titles not only contributes to audience engagement but also ensures that the content is deemed appropriate and enjoyable by a wide range of viewers. Positive sentiments are generally associated with content that is inclusive, inspiring, and family-friendly.
- Neutral Sentiment:
  - Videos with neutral sentiments in their titles may not evoke strong emotions. They are often informative or present factual information without a distinct positive or negative tone. These videos are typically suitable for a broad audience, appealing to viewers seeking unbiased and straightforward content.
- Negative Sentiment:
  - Videos categorized with negative sentiment in their titles may contain content that is controversial, critical, or potentially upsetting. While negative sentiments can be powerful

for expressing opinions or addressing serious topics, they may limit the appropriateness of the content for a broader audience.

In short, by understanding the sentiments in video titles, content creators can effectively tailor their material to reach the right audience and foster a positive and inclusive online community.

About Dataset

The dataset is daily updated by "Rishav Sharma". The dataset is available on Kaggle. (https://www.kaggle.com/datasets/rsrishav/youtube-trending-video-dataset)

The sentiment analysis is performed on a comprehensive dataset comprising YouTube video information from multiple countries. This dataset serves as a daily record of the top trending YouTube videos, updated regularly. According to Variety magazine, YouTube determines the top-trending videos based on various factors, including user interactions such as views, shares, comments, and likes. It's essential to note that these are not necessarily the most-viewed videos overall for the calendar year.

The dataset includes several months of data on daily trending YouTube videos across different regions, including India (IN), the United States (US), Great Britain (GB), Germany (DE), Canada (CA), France (FR), Russia (RU), Brazil (BR), Mexico (MX), South Korea (KR), and Japan (JP). Each region's data is stored in a separate file, providing information on video titles, channel titles, publish time, tags, views, likes and dislikes, description, comment count, and categoryId.

The categoryId field varies between regions, and to retrieve the categories for a specific video, one can refer to the associated JSON files included for each region in the dataset. The data was collected using the YouTube API, and this dataset represents a structurally improved version of the original.

Current State of the Art

The existing landscape of sentiment analysis on YouTube videos reveals a predominant focus on data analysis, with limited research specifically addressing sentiment analysis or natural language processing. Notably, there is a scarcity of fine-tuned BERT models, such as "nlptown/bert-base-multilingual-uncased-sentiment," tailored for sentiment classification tasks across diverse languages.

While exploring related work, a few research papers were identified, contributing to the sentiment analysis domain:

- Sentiment Analysis for Youtube Videos with User Comments: Review
  - Link: https://ieeexplore.ieee.org/abstract/document/9396049
  - This paper primarily concentrates on sentiment analysis within the United States, covering English and Indonesian languages. However, its scalability is limited due to the focus on specific languages within a single country.
- Discovering popular and persistent tags from YouTube trending video big dataset
  - Link: https://link.springer.com/article/10.1007/s11042-023-16019-z

– The second research paper evaluates sentiment analysis on YouTube trending videos, specifically in the United States for the year 2021. The approach is tailored to a particular country and time period, lacking generalization across multiple countries and years.

This project aims to improve YouTube sentiment analysis by integrating -

- more data
- data of many years (from last couple of years to the data which is generated today)
- data from 11 countries varying in languages of video titles/ tags/ comments

Methods

Data Collection and Preprocessing

The initial dataset comprises 11 .csv files, each representing a country's daily trending videos. Spanning from 08/06/2020 to 10/30/2023, one file contains over 235,000 records. Utilizing pathlib, the data is consolidated into a dataframe, with a new 'country' column derived from the initials in the .csv filenames. The size of combined dataset is 2,552,859 rows and 17 columns.

Date-related operations are facilitated by creating two additional columns, 'publish_date' and 'publish_time,' extracted from the 'publishedAt' column. The 'trending_date' column is split to isolate the date information. To gauge the duration a video trends, a 'days_of_trending' column is computed as the difference between 'trending_date' and 'publish_date.'

Number of days the video was trending are counted with the following syntax:
data['days_of_trending'] = (data["trending_date"] - data["publish_date"]).dt.days

I believe that the video is from a country where it is in trending for the most days. For example, if there is a video of president of United States, then the video should have the most trending days in United States, and the video is from the content creator o United States. So, I sort days_of_trending column in descending order, and deleted the duplicate rows (the same video may be trending in India, but has the least trending days) which have same title.
On doing the analysis on days_of_trending column, it reveals that 50% of videos trend for less than 5 days, with a maximum trending duration of 37 days.

Further exploration uncovers a correlation between view count and likes, while dislikes and comment count exhibit no significant correlation due to many videos disabling these features.

There are additional .json files with the dataset. These files contain the information about the video category. There are total 11 files, one for each country. These files have categoryId and the name of that category. These categoryId and name of the category may vary according to country. These json files are incorporated into dataset and mapped the category with categoryId. Analysis on category shows that entertainment category videos are trended the most over the years.

This probelm is from Kaggle and there are some YouTube video analysis codes available on Kaggle. Few codes are available with data analysis and data preprocessing steps.

Language Detection

Language diversity in video titles necessitates language detection using the NLTK library to remove stopwords. Considering the varied linguistic landscape, stopwords from multiple languages (Portuguese, German, English, Hinglish, French, Russian, and Spanish) are employed. The dataset created with the help of stopwords will later be used in the BERT model for model training.

The BERT model(nlptown/bert-base-multilingual-uncased-sentiment) supports sentiment analysis in English, Spanish, German, and French language. To align with the capabilities of the chosen BERT model, data from the United States, France, Mexico, and Germany is selected.

There can be a case when the video of one country is in trending in another country. Langid library is used to detect the language of video title, and kept only those data where the video title's detected language is English, French, German, and Spanish. Langid library detected many other languages in video titles, but only the above mentioned languages are kept because the BERT model only supports those languages.

After that, to reduce the data and make data more robust, a condition was applied that matches detected language and country column.
For example, if detected language is French, then country must be France.

The syntax for that is as follows:
condition = ((data['detected_language'] == 'en') & (data['country'] == 'US')) | ((data['detected_language'] == 'de') & (data['country'] == 'DE')) | ((data['detected_language'] == 'fr') & (data['country'] == 'FR'))| ((data['detected_language'] == 'es') & (data['country'] == 'MX'))

After these preprocessing steps, the dataset is streamlined to 87,428 rows, facilitating more focused and efficient analysis compared to the initial expansive dataset.Out of 87,428 rows, 67,000 rows are used as train data and remaining rows are used test data

Sentiment Classification

A sentiment analysis pipeline, instantiated using the Hugging Face transformers library, is employed to utilize a pre-trained BERT model (nlptown/bert-base-multilingual-uncased-sentiment). This model is specifically designed for sentiment analysis tasks across multiple languages.

The sentiment analysis pipeline is utilized to classify sentiments in YouTube video titles. For each title, the model outputs a sentiment label and a corresponding confidence score.

Before feeding the titles into the sentiment analysis model, several preprocessing steps are applied to enhance the model's performance. These include:

- Lowercasing all titles for uniformity.
- Removing punctuation to focus on the textual content.
- Cleaning titles by eliminating non-alphanumeric tokens, stopwords, and any entries resembling monetary amounts.

To facilitate the integration of textual data into a neural network, a Tokenizer is created using the TensorFlow Tokenizer class. The titles are tokenized and sequences are padded to a fixed length (max_length) using TensorFlow's pad_sequences function. This ensures uniform input dimensions for the subsequent neural network model for both the train and test dataset.

Model Training and Evaluation

A custom neural network model is constructed for sentiment classification. This model incorporates a pre-trained BERT model as the initial layer, with its layers frozen to retain the learned representations. The BERT output is followed by a dense layer of 256 neurons with rectified linear unit (ReLU) activation and a final dense layer of 3 neurons with a softmax activation function to classify sentiments into three categories: positive, negative, and neutral.

After that, the dataset is split into training (X_train, y_train) and validation (X_val, y_val) sets using a test size of 20%. The sentiment labels are encoded using a LabelEncoder to convert categorical labels into numerical values.

To train the sentiment analysis model, Stochastic Gradient Descent (SGD) optimizer is utilized. The model is compiled using the sparse categorical cross-entropy loss function. The compilation also specified the inclusion of accuracy as a metric for evaluating model performance.

The model is trained for a total of five epochs, with each epoch comprising batches of 32 samples.Throughout training, the ModelCheckpoint callback is employed to monitor the validation accuracy and save the model weights when improvements are detected. This precautionary measure ensures that the best-performing model is retained, preventing potential performance degradation during training.

Experiments and Results

The experimental setup involves fine-tuning a pre-trained BERT model on a labeled sentiment dataset for YouTube video titles. The labeled dataset is split into training and testing sets, and sentiment labels are encoded for compatibility with the model.

During training, the model's performance is monitored using a validation set, and the best weights are saved using the ModelCheckpoint callback. The training curve is plotted to visualize the accuracy trends over epochs.

The experiment is run for a total of ten epochs, with the initial five epochs showing improvement in accuracy. Subsequently, the best-performing model is loaded for further evaluation.

Model Testing

The final stage involves testing the model on an independent dataset, the 'test_data,' to evaluate its generalization to unseen instances. The 'best_model' is loaded using the 'load_model' function. The model is then evaluated on the test set using the 'evaluate' method, and the test accuracy is printed to assess its performance on previously unseen data. The model got the accuracy of around 65%.

Conclusion

In conclusion, this project explores sentiment analysis on YouTube videos, employing a comprehensive approach with a diverse dataset spanning 11 countries and multiple years. The meticulous collection and preprocessing of data have resulted in a refined dataset, providing a solid foundation for in-depth exploration of sentiment trends on the platform. Unlike existing research that often focuses on specific languages or individual countries, this project strives for broad generalization, offering valuable insights into sentiment dynamics across diverse linguistic landscapes and cultural contexts.

The integration of a fine-tuned BERT-based sentiment analysis model, such as "nlptown/bert-base-multilingual-uncased-sentiment," further elevates the project's analytical capabilities. Leveraging state-of-the-art language models, the sentiment classification model is adept at discerning positive, negative, and neutral sentiments in video titles across multiple languages. With refined methods and an expansive dataset, this project stands as a significant contribution to the field, opening avenues for a more comprehensive understanding of sentiment patterns in the dynamic realm of YouTube content.

Citations for data analysis and pre-processing:

- https://www.kaggle.com/code/venky73/youtube-trending-videos-analysis
- https://www.kaggle.com/code/singatharun/recommendation-for-youtube-content-creators
- https://www.kaggle.com/code/jovianreynaldo/youtube-trend-classification
- https://www.kaggle.com/code/ericjohndungdung/youtube-india-2022-year-trends
- https://www.kaggle.com/code/mohitmanjaria/youtube-trending-analysis-rf-and-xgboost/notebook
- https://www.kaggle.com/code/kooose/categorical-analysis-and-predict-lstm
- https://www.kaggle.com/code/wardalaknaoui1/sentiment-analysis-usa-step-by-step-emojis-tags
- https://www.kaggle.com/code/italomarcelo/eda-prediction