Classification of YouTube data based on Sentiment Analysis

Rupesh Bamane[1], Mohanish Raul[2], Ajay Vadkar[3]

[1]Rupesh Bamane, Dept of Information and Technology, Atharva College of Engineering, Maharashtra, India
[2]Mohanish Raul, Dept of Information and Technology, Atharva College of Engineering, Maharashtra, India
[3]Ajay Vadkar, Dept of Information and Technology, Atharva College of Engineering, Maharashtra, India

---***---

Abstract - Corporate companies are using social media for improving their business, the data mining and analysis are very important in these days. Thus Interaction and Review are established with the customer and the concept, characteristics and need for Big Data and different offerings available in the market to explore unstructured large data. The paper deals with the analysis of YouTube Data. The study is done using users Sentiments features such as Views, opinion, Likes and Dislikes.

Key Words: YouTube Data, Sentiments Features, Views, Comment, Dislikes.

1. INTRODUCTION

Classification of youtube data totally based on sentiment analysis will be able to analyse the data that associate with the youtube. Project will be able to analyse data related to youtube based on likes, views, comments as well as it will give country wise data analysis. Youtube sentiment analysis will have full videos data of country like United States, Great Britain. Youtube sentiment analysis will be able to analyse research community continuously showed keen interest in analyzing and exploiting the rich content shared on YouTube. Based on youtube data like getting the data of popular video of a particular channel project give analytics like how many likes, comments and views or dislike on this video. Project gives response in various formats like text, graphs, etc.

1.1 NEED

In recent year's online social media like Facebook, Twitter, YouTube and Google+ make the space for millions of users to share their information and opinion with each other. With the blistering favor, these sites have become a source of massive amount of real time data of videos, images etc. Among them, YouTube is one of the world's largest video sharing platforms, where videos are uploading continuously by the millions of users (companies, private individual, etc.). YouTube has emerged as a cyclopedic and accessible compilation of video information source on the web.

1.2 Application and Scope

Application:

A user can upload dataset related to required information and System will give graphical representation of the uploaded unique dataset.

SCOPE:

In order to escalate the user's communication it allows users to express their opinion by rating the viewed objects (by clicking on the like/dislike buttons) and interacting with the other community members (via the comments feature). These activities (like / dislike /number of views) of the users can serve as a global indicator of immense or popularity for a demanding video. These Metadata serve the purpose of helping the community to filter relevant opinions more efficiently. When we search for a specific video through some keyword on specific object, the most popular video comes (which are rated based on actual views/likes by the users) first in search panel based on that given unique keywords.

1.3 AIM

The Classification of YouTube data based on Sentiment analysis aims to provide efficient and accurate graphical representation of like, views from the dataset using Latent Semantic Analysis (LSA).

1.4 Problem Statement

To overcome with these problems, we have designed a System for Youtuber or content creators (or users) to give accurate graphical representation of like, views and sentiment analysis of comments from the uploaded unique dataset of videos to let them simplify their work and increase their productivity.

2. Literature surveyed

From this journal we calculated that Video-sharing websites such as YouTube have become a channel for spreading bigotry and being used as an Internet based distribution platform for like-minded people to interact, publicize and share their ideologies. Due to low publishing hudle, websites such as YouTube contains a large database of user generated content (UGC) in the form of videos and textual comments which are malicious and racist. Online bigotry and hate content can have a negative brunt on society and the prevalence of such easily accessible content is thus a major concern to the people, government and law enforcement agencies. Solutions to counter cyberattack-crime related to the promotion of hate and radicalization on the Internet is an area that has recently attracted a lot of research attention. Hence from this paper we studied that to retrieve hate and extremist videos, users and communities from YouTube. These proposed system able to bootstrap from 60 (seed-list) to 158 (true positive) users in two emphasis. The system was able to search 98 users automatically with a rigor of 88%. The proposed approach can discover central, and authoritative users and videos as well as hidden communities using social network analysis.

Opinion mining is a form of opinion analysis towards a pattern or mood of a person or a certain subject, these things are often called sentiment. In order to filter every statement or opinion sentiment from the public, the most and easy publishing media that is being used is the internet. Many comments on Ahok's remarks have been made on YouTube. The process of seeking or tracing the natural language to find patterns or moods of society against certain products, people or topics is called Sentiment Analysis. Sentiment analysis is also often referred to like the opinion of mining. Hence we concluded from this paper that Support Vector Machine (SVM) is being used to view the performance sentiment analysis of Ahok. There are four cases that has been attended namely, Data Comments, Pre-Processing, Tokenizing and regulate Sentiment with Lexicon Based. Davious the percentage density in this research had used Lexicon Based and Confusion Matrix to know the result of the weighting percentage of analysis to SVM. Sentiment reasoning can be used to find out how far the performance of Ahok based on the results gained from netizen's comment on YouTube. The result of a classification of weighted values according to the Support Vector Machine (SVM) method has brings us to the conclusion that the value of True Positive rate is 91.1% based on the comments taken from 2015 until 2016. For the further development of this research, further researchers need to take into account that data recording should be very high in number to achieve the accuracy of the results and conclusions on the opinion mining analysis.

3. Requirement Analysis

User:

Youtuber whos channel is on youtube is the user of this system. Youtuber enters data on websites upload data section and gives to the user.

Webpage: localhost

This system works using LSA technologies. This web pages gives inputs from the user end and processes it to find out the proper meaning of data. Then it gives responses in various formats such as text, graphs, tables, etc.

Database:

Database is very important to make this system work. The database consists of all video database as well as comment datasets. Without the database, this system will work only for login window. Database makes it work for specific queries. Databases design in SQL format.

4. Implementation

4.1 Comment Collection and Pre-processing:

The aim of this section is acquisition of opinion on a selected YouTube video. In order to address this task a focused crawler is implemented. According to the video URL, it quotation comment (up to 1000) of that video using web API through HTTP GET method. But, the extracted comments are heterogeneous in terms of languages and various notions used by the users. Therefore, we carried out some pre-processing on these unstructured comments to generate the data sets.

4.2 Generating Data Sets:

For each evaluating video two datasets are made according to the proposed approach. To make datasets, in the processed text MySQL stop word is applied to remove all the stop words and then convert all the words into their singular form and thus make dataset 1st. Next, for Dataset 2nd all the adjectives of the comments are gathered.

4.3 Sentiment Measure:

Project use SentiStrength3 thesaurus in both the dataset to assess the overall tendency of user comments. SentiStrength is a sentiment lexicon analysis classifier which appraisal the tenacity of positive and negative sentiment of the comment words. It address two sentiment strengths:- 1 st(not negative) to -5th (extremely negative) and 1st (not positive) to 5th (intensely positive). If a word within a sentence got <1 rating, the classifier select it as a negative word and if got >=1 rating, then classifier selects it as a positive word.

4.4 Video Rating:

In this section the statistical comparison of the standard deviation (SD) values is conducted (both positive and negative distributions) across both dataset. After calibrating the sentiment rate of both data set the standard deviation technique is applied.

4.5 Sentiment Analysis

We follow the standard sentiment classification approach. We use the Naive Bayes classification technique for sentiment analysis.The comments we collected for each keyword is used as the test data for classification. The Naive Bayes classifier is trained on the opinion from the training set and is then used to determine the overall sentiment for each comment in the test set. A review or comment is considered as an independent words (i.e., the ordering of the words is not considered). The positive and negative opinion in the trained dataset are stored in two separate dictionaries, which we refer to as positive dictionary (positive opinion) and negative dictionary (negative opinion). For each opinion, the polarity/sentiment of each word is determined by calculating the number of times the word appears in the positive and negative dictionaries. For each word, the positive antinomy is number of times the word appears in the positive dictionary divided by the total number times it appears in both the positive and the negative dictionaries. Similarly, the negative polarity is the number of times the word appears in the negative dictionary divided by the total number of times it appears in both dictionaries.

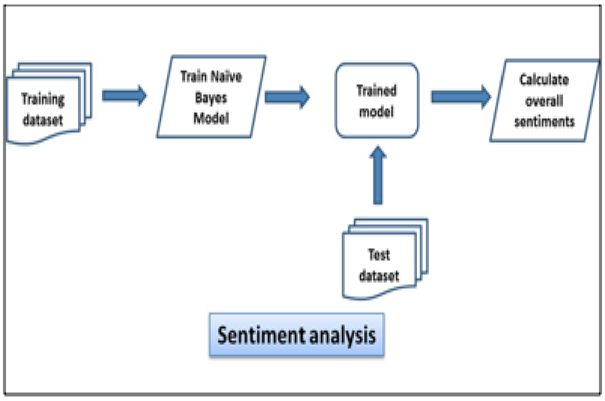


Fig.6.5.1 Standard Sentiment Classification Approach

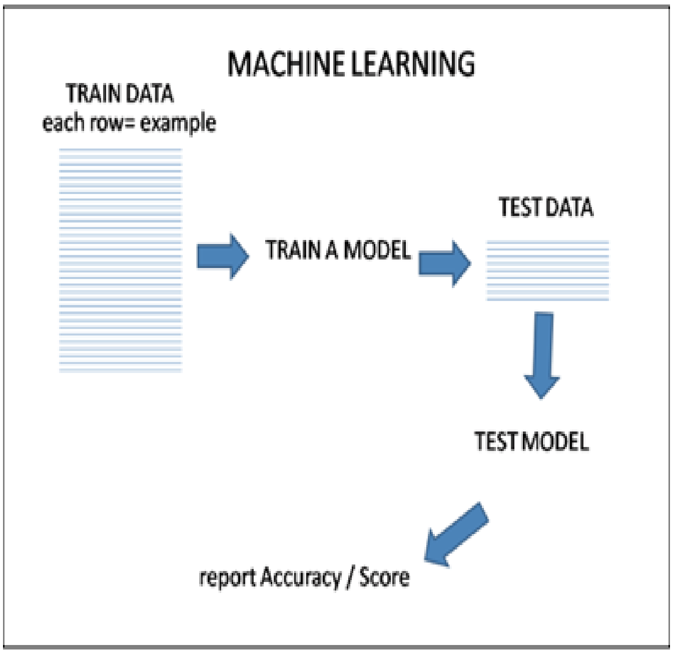
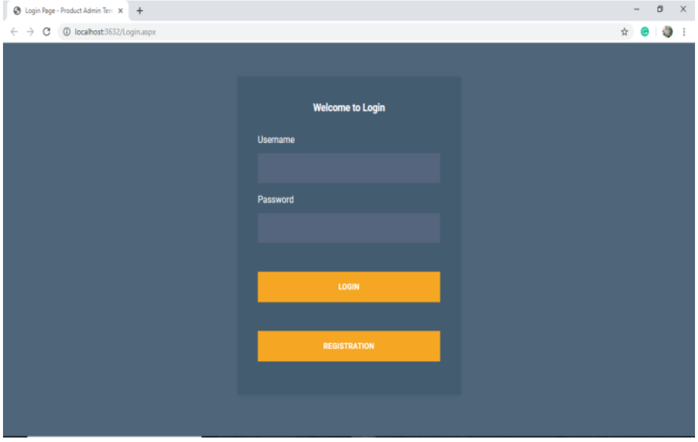


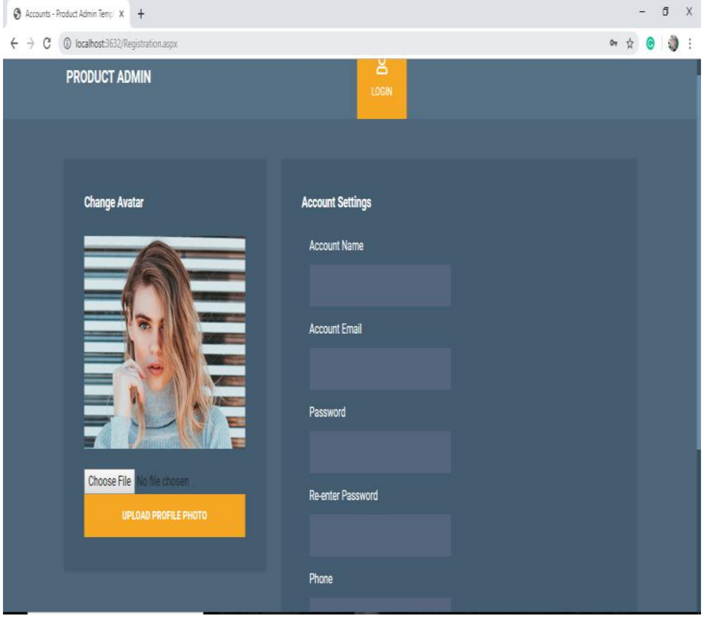
Fig 6.5.2 Overview of the entire sentiment analysis process

5. Results:

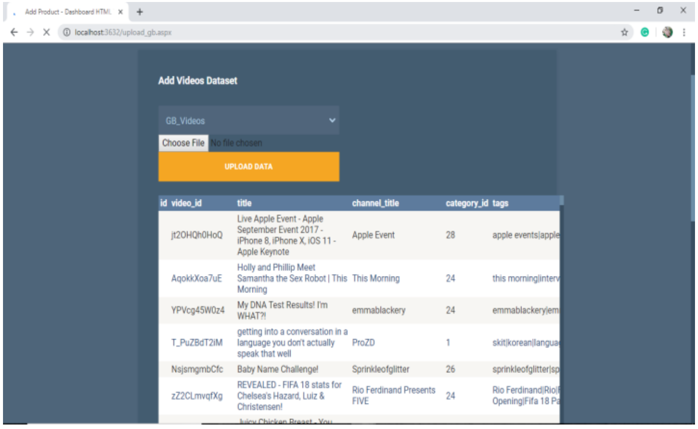
Login Page



User Registration

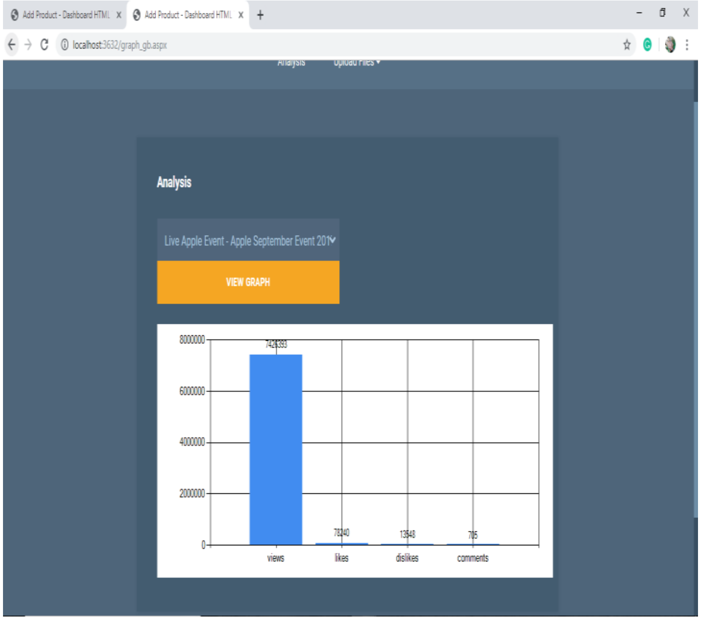


Data can be viewed

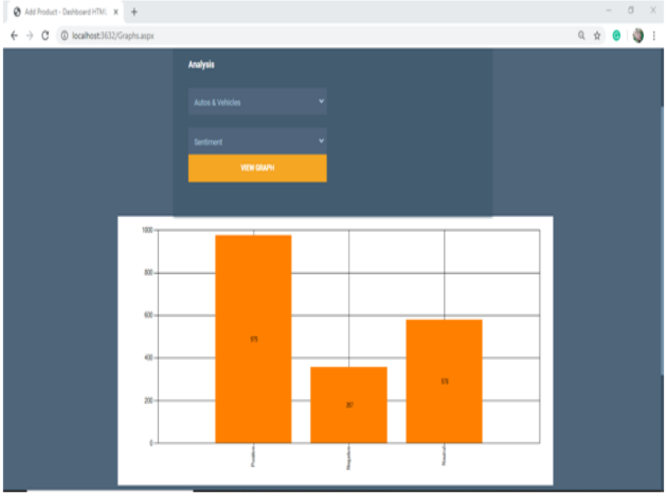


id	video_id	title	channel_title	category_id	tags
	jt2OHQh0HoQ	Live Apple Event - Apple September Event 2017 - iPhone 8, iPhone X, iOS 11 - Apple Keynote	Apple Event	28	apple events|apple
	AqokkXoa7uE	Holly and Phillip Meet Samantha the Sex Robot | This Morning	This Morning	24	this morning|interv
	YPVcg45W0z4	My DNA Test Results! I'm WHAT?!	emmablackery	24	emmablackery|em
	T_PuZBdT2iM	getting into a conversation in a language you don't actually speak that well	ProZD	1	skit|korean|langua
	NsjsmgmbCfc	Baby Name Challenge!	Sprinkleofglitter	26	sprinkleofglitter|sp
	zZ2CLmvqfXg	REVEALED - FIFA 18 stats for Chelsea's Hazard, Luiz & Christensen!	Rio Ferdinand Presents FIVE	24	Rio Ferdinand|Rio| Opening|Fifa 18 Pa
		Juicy Chicken Breast - You			

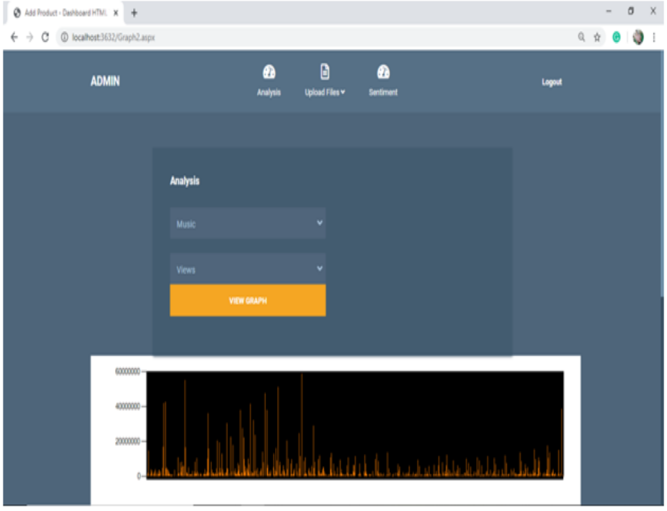
Video analysing


Add Product - Dashboard HTML
Add Product - Dashboard HTML
localhost:3632/graph_gb.aspx
Analysis
Upload Files
Analysis
Live Apple Event - Apple September Event 201
VIEW GRAPH
8000000
6000000
4000000
2000000
0
7426393
78240
13548
705
views
likes
dislikes
comments

Polarity Analysis



Graphical Analysis of Views



6. Conclusion

This project will definitely simplify the work of YouTube and content creator. By showing them their or others youtubers video analysis based on their views likes and comments from dataset what kinds of positive and negative comment has occurred on that video so user analyse them and create better content or they get motivated for uploading more videos.

7. References

[1] A. Severyn, A. Moschitti, O. Uryupina, B. Plank and K. Filippova,"Multi-lingual opinion mining on youtube," Information Processing & Management, 52(1), 2016, pp. 46-60.

[2] S. Chelaru, C. Orellana-Rodriguez and I. S. Altingovde, "How useful is social feedback for learning to rank YouTube videos?" In World Wide Web, 17(5), 2013, pp. 1-29.

[3] P. Schultes, V. Dorner and F. Lehner, "Leave a Comment! An In- Depth Analysis of User Comments on YouTube," Wirtschaftsinformatik, 2013, pp. 659-673.

[4] S. Siersdorfer, S. Chelaru, J. S. Pedro, I. S. Altingovde and W. Nejdl, "Analyzing and mining comments and comment ratings on the social web," ACM Transactions on the Web (TWEB), 8(3), 2014, pp. 1-39.