Sentiment Analysis on YouTube

Sameera Turupu
Master's in Computer Science
Georgia State University
Atlanta, Georgia
sturupu1@student.gsu.edu

Aditya Bhamidipati
Master's in Computer Science
Georgia State University
Atlanta, Georgia
vbhamidipati1@student.gsu.edu

ABSTRACT

Sentiment analysis or opinion mining is the field of study related to analyze opinions, sentiments, evaluations, attitudes, and emotions of users which they express on social media and other online resources. YouTube is one of the comprehensive video information sources on the web where the video is uploading continuously in real-time. It is one of the most popular sites in social media, where users interact with sharing, commenting, tagging and rating (like/dislikes) videos. Generally, the quality, relevance, and popularity of the video is maintained based on this rating. Sometimes irrelevant and low-quality videos ranked higher in the search result due to the number of views or likes, which seems untenable. Tags play an important role especially for YouTubers as they earn money from advertisers based on their views. They often ponder with the question "What type of videos should I make in order to get more user views?" and more specifically what should be the title and tags to attract users. To minimize these issues, we present a sentimental analysis approach to user comments and tags using Natural Language Processing. This project offers a novel approach to analyze opinions posted by users about a video and provides insights on most frequently used tags.

Keywords

YouTube; Sentiment; Analysis; Comments; Mining; Views; Likes; Dislikes; Tags; Natural Language Processing

INTRODUCTION

Social media and streams, such as Twitter, Facebook, YouTube, and Google+, make the space for millions of users to share their information and opinion. With the rapidly increasing popularity, these sites have become a source of massive amounts of real-time data of videos, images, etc., that can dramatically affect the reputation of a person or an organization. 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 persons, etc.). YouTube has emerged as a comprehensive and accessible compilation of video information sources on the web. It is a unique environment with many facets, such as multi-modal, multilingual, multi-domain, and multi-cultural.

The versatility of variety and attractive shared content draw widespread attention. Therefore, the importance of YouTube is successively increasing for the industry and research community day by day. YouTube was ranked as the second most popular site by Alexa Internet, a web traffic analysis company in Dec 2016. According to a recent study, YouTube accounts for 20% of Web traffic and 10% of total Internet traffic. To increase the user's interaction, YouTube provides many social mechanisms to judge user opinions and views about a video through voting, rating, favorites, sharing and negative comments, etc. These activities (like / dislike /number of views) of the users can serve as a global indicator of quality or popularity for a video. Moreover, these Meta data (like/dislike/number of views) serve the purpose of helping the community to filter relevant opinions more efficiently.

For instance, consider a typical comment on a YouTube review video about an Apple Mac book pro vs Microsoft Surface laptop 3:

this girl really puts a negative spin on the Microsoft Surface Line up, and I 'm not sure why, this seems crazy fast, and I 'm not entirely sure why her pinch to zoom on touch is different from all the other surface reviews

The comment contains a product name surface 3 laptop and some negative expressions, thus, a bag-of-words model would derive a negative polarity for this product. Similarly, we can find the overall duality of the video by calculating the polarity of all the comments and then comparing them with the likes to evaluate the relevance.

YouTube is one of the most popular platforms for making money online. The biggest challenge that every youtuber face is "which videos grab more user's attention(views)?" more specifically, what should be the video content so that the video can obtain more user views. Search results also focus on title and tags, so analysis on tags will help increase the popularity of a video.

RELATED WORK

An analysis of the social video sharing platform YouTube reveals a high amount of community feedback through comments for published videos as well as through meta ratings for these comments. In-depth study of commenting and comment rating behavior is necessary to analyze dependencies between comments, views, comment ratings and topic categories [6]. The influence of sentiment expressed in comments on the ratings for these comments using the SentiWordNet thesaurus [7], a lexical WordNet-based resource containing sentiment annotations, gives an idea on polarity of a video. A vast amount of social feedback expressed via ratings (i.e., likes and dislikes) and comments is available for the multimedia content shared through Web 2.0 platforms. However, the

potential of such social features associated with shared content remains unexplored in the context of information retrieval [2]. We investigate the effectiveness of each such feature for video retrieval and the correlation between the features. Finally, we investigate the impact of the social features on the video retrieval effectiveness using statistical approaches.

Previously an unsupervised lexicon-based approach [5] is used to detect the sentiment polarity of user comments in YouTube. Polarity detection in social media content is challenging not only because of the existing limitations in current sentiment dictionaries but also due to the informal linguistic styles used by users. Present dictionaries fail to capture the sentiments of community-created terms. To address the challenge, we adopted a data-driven approach and prepared a social media specific list of terms and phrases expressing user sentiments and opinions by preprocessing the data. To predict community acceptance for videos (Positive or Negative), evaluation of different Machine learning classifiers for the estimation of ratings for these comments is necessary [6]. The results of a large-scale evaluations suggest the community acceptance of a video.

On the Internet, the term "tag" refers to keywords assigned to an article, image, or video. With the rapid development of social sharing websites (i.e., Flickr, Picasa, and YouTube), the tags help organize, browse and search relevant items within these massive multimedia collections [13]. Given the majority of tag types, YouTubers should get an idea of what type of content to prepare to attract its users.

MOTIVATION

It was very much an elementary study, but the desire was to try to discover more formally how accurate this belief was, that is, and how could the level of relevance of video be measured? This raises the importance of the automatic extraction of sentiments and opinions expressed in social media. While sentiment analysis for more current data has recently attracted much attention from both industry and academia, the lack of manually annotated data makes these studies only partially useful for social media and streams. As YouTube has an API, this means that some information from the website can be legally accessed using a developer account. The goal was to extract comments associated with particular video postings and then investigate how relevant the video is to the comment. The added value from this could facilitate further sociological investigations into a modern phenomenon and be useful for a proper search.

PROBLEM STATEMENT

When we search for a specific video through some keyword on one particular topic, the most popular video comes (which are rated based on views/likes by the users) first in search panel based on that given keywords. Therefore, sometimes some problematic issues arise in searching, such as inconsistency, irrelevancy, etc. For example, when we do a query like "Avengers: Infinity War." For that search, none of the results of videos were actually of that movie. Most of the results for this query are inconsistent. Among the result, some are movie's cut scenes, and some are movie's specific clip like action, romance, and some are about its trailer. However, there is nothing actually what we look for. This kind of scenario occurs in YouTube frequently. Until we run the video, we could not understand. This situation comes because of the number of views and likes of those videos. Therefore, to find the perfect and relevant video, an effective and efficient process seems conceivable, which would not depend on only those metadata (like/dislike/number of views).

Dataset Collection

The data is collected from (up to) 200 listed trending YouTube videos every day in the US and the UK. The dataset includes data gathered from videos on YouTube that are contained within the trending category each day. There are two kinds of data files, one includes comments and one includes video statistics. They are linked by the unique video_id field.

The headers in the video file are: video_id (Common id field to both comment and video csv files), title, channel_title, category_id (Can be looked up using the included JSON files, but varies per region so use the appropriate JSON file for the CSV file's country), tags (Separated by | character, [none] is displayed if there are no tags), views, likes, dislikes, thumbnail_link, date (Formatted like so: [day].[month]) The headers in the comments file are: video_id (Common id field to both comment and video csv files), comment_text, likes, replies

Dataset: https://www.kaggle.com/datasnaek/youtube.

video_id	comment_text	likes	replies
XpVt6Z1Gjjo	Logan Paul it's yo big day â€¼ï¸â€¼ï¸â€¼ï¸	4	0
XpVt6Z1Gjjo	wing you from the start of your vine channel and have see	3	0
XpVt6Z1Gjjo	Say hi to Kong and maverick for me	3	0
XpVt6Z1Gjjo	MY FAN . attendance	3	0
XpVt6Z1Gjjo	trending ðŸ˜‰	3	0

Figure 1: Input Structure of Comment file

video_id	title	channel_title	category_id	tags	views	likes	dislikes	comment_total	thumbnail_link	date
XpVt6Z1GjjoGAN PAU	Logan Paul Vlog		24	button|10	4394029	320053	5931	46245	https://img.com/vi/XpVt6Z1Gjjo/c	13.09
K4wEI5zhHEoducing iP	Apple		28	|augmente	7860119	185853	26679	0	https://img.com/vi/K4wEI5zhHB0/	13.09
cLdxuaxaQwlty Respons	PewDiePie		22	[none]	5845909	576597	39774	170708	https://img.com/vi/cLdxuaxaQwc/	13.09
WYYvHb03EiPhone X fir	The Verge		28	cs|iPhone	2642103	24975	4542	12829	https://img.com/vi/WYYvHb03Eog/	13.09
sjlHnJvXdQone X (parc	jacksfilms		23	phone|iph	1168130	96666	568	6666	https://img.com/vi/sjlHnJvXdQs/c	13.09

Figure 2: Input Structure of Video file

Output description

The output of algorithm is an emoticon that suggests the polarity of a video based on the comments and likes separately so as to compare the correctness of user acceptance. Also, a prediction model is built for the program that can predict a video polarity based on comment text. Additionally, analysis on tags outputs the top 30 tags for the most trending content on YouTube.

PROJECT DESCRIPTION

This section describes the NLP-based methodology of sentiment analysis on user's comment in order to retrieve the most relevant and perfect YouTube videos. The proposed process works in four steps as shown in Figure. 3

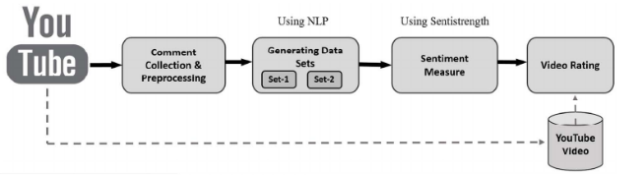

Figure 3: Overall work process of sentiment analysis on user comments

First, comment collection is done from the dataset repository and exploratory data analysis is performed followed by pre-processing module extracts data (comments) from the specific YouTube video and do some language pre-processing to prepare for the next process. Then, the processed text go through a NLP-based methods to generate data sets. Next, apply the sentiment classifiers on the data sets to calculate the positivity and negativity scores. Finally, visualization is performed. In detail the methodologies are explained below.

Exploratory Data Analysis

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns,to spot anomalies,to test hypothesis and to check assumptions with the help of summary statistics and graphical representations. It is a good practice to understand the data first and try to gather as many insights from it. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task.

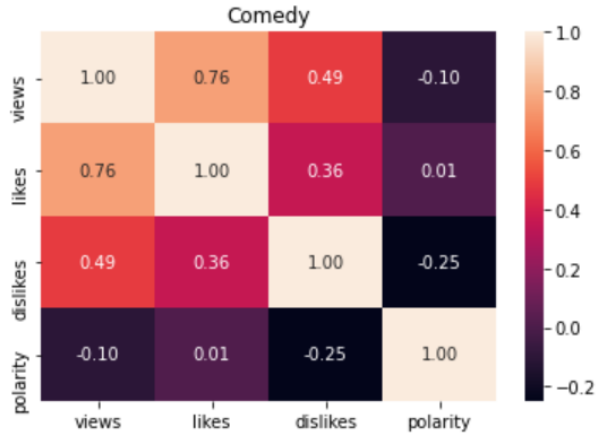

Figure 4: Correlation matrix between metadata

We have built a correlation matrix for each category like comedy, entertainment etc showing the correlation between likes, views, dislikes and polarity of the comments. The values of the correlation coefficient can range from -1 to +1. The closer it is to +1 or -1, the more closely are the two variables related. The positive sign signifies the direction of the correlation i.e. if one of the variables increases, the other variable is also supposed to increase. Here views and likes are highly correlated whereas dislikes and polarity are least correlated which means the number of dislikes is not proportional to comments polarity for comedy category.

Figure 5: Word cloud for positive comments

Figure 5 is the word cloud for positive comment sentences as you can see the words lol, think etc are most used in the positive comments.

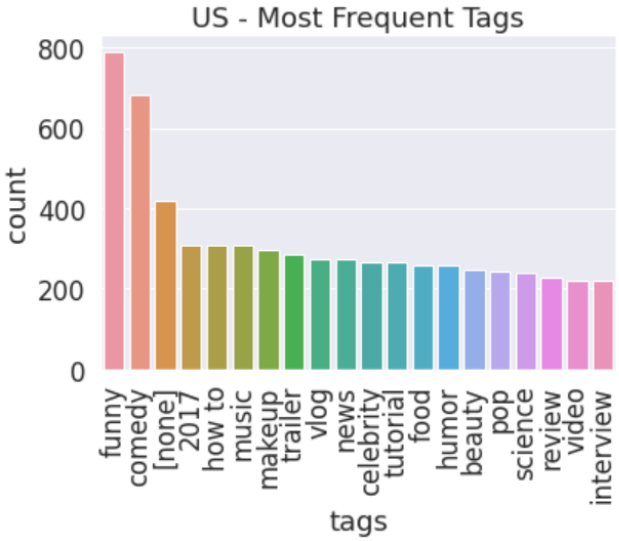

Figure 6: Analysis on most frequent tags

Figure 6 is an analysis on most frequent tags on US YouTube videos. As you can see, funny and comedy tags are highly used. Review and interview are least used tags on US Youtube videos.

Preprosessing

Remove all the expressions which are irrelevant for the proposed methodology like link (www.abc.com, www.pqr.com etc.), numbers (192, 500 etc.) and special characters ("*","/","!","","?","#","&","$"), emoticons (“:), :D, <3, ;) etc.”). Remove all the punctuation marks such as period (“.”), space ("-"), commas (","), semicolon (“;”), hash (“-“) etc. Remove stop words from the comment texts as they don’t add any weight for processing. A stop word is a commonly used word (such as “the”, “a”, “an”, “in”) that a search engine has been programmed to ignore, both when indexing entries for searching and when retrieving them as the result of a search query.

The later stages are tokenization and lemmatization. Tokenization is the process of tokenizing or splitting a string, text into a list of tokens. One can think of token as parts like a word is a token in a sentence, and a sentence is a token

in a paragraph. Eg., “This is a sample” -> [“This”, “is”, “a”, “sample”]. Lemmatization takes into consideration the morphological analysis of the words. To do so, it is necessary to have detailed dictionaries which the algorithm can look through to link the form back to its lemma. We have used WordNet lemmatizer. Eg., “singing” -> “sing”

ALGORITHM DESIGN

Architecture

The data extracted from youtube api and kaggle data set are saved to local file system in KVM. Data from local system is loaded to spark engine using pyspark API (python). Spark sends data and tasks i.e map and reduce to worker nodes to do the parallel processing. Spark sends output back to the local system and then output is fed to matplotlib and tableau for visualizations and the results are used for decision making.

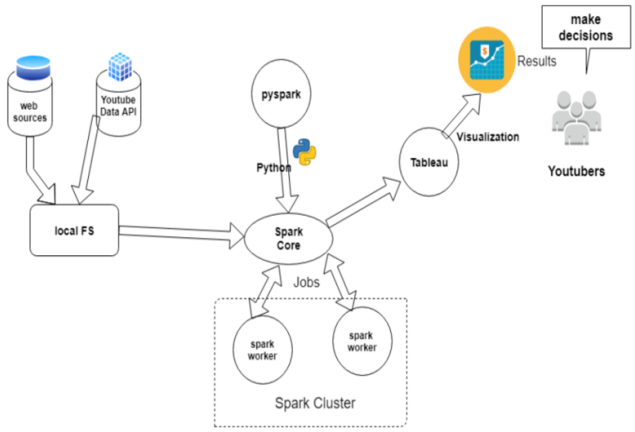

Figure 7: Project Architecture

Data Pipeline

Data collected from different sources are saved locally and fed to spark. Spark transmits data into RDDs to its worker nodes. All the worker nodes execute their tasks and send the sorted output back to the local FS. Outputs are visualized in tableau for decision making.

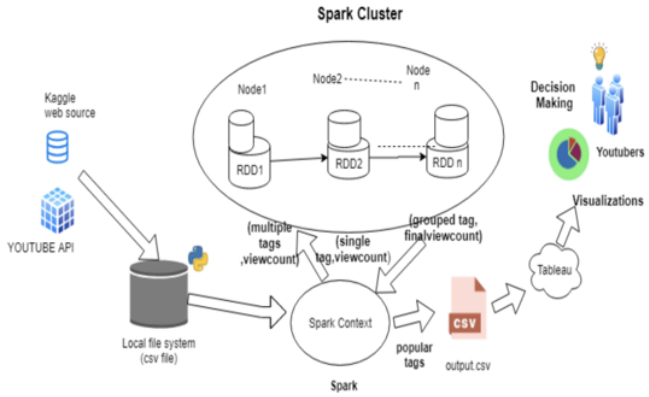

Figure 8: Data pipeline for Tag analysis

Machine Learning Algorithms

Textblob: is a python library for processing textual data (comments text). It provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more. It’s a sentiment lexicon (in the form of an XML file) which it leverages to give both polarity and subjectivity scores. The polarity score is a float within the range [-1.0, 1.0]. The subjectivity is a float within the range [0.0, 1.0] 0.0 is very objective and 1.0 is very subjective. Text classification occurs using NaiveBayes Classification in TextBlob.

Logistic Regression: We permormed multiclass logistic regression on analysed data using spark ML. It has few preprocessing steps. Firstly, the StringIndexer class performs label encoding and must be applied before the OneHotEncoder which in turn converts categorical variables into a form that could be provided to ML algorithms to do a better job in prediction (The categorical value represents the numerical value of the entry in the dataset). The VectorAssembler class takes multiple columns as input and outputs a single column whose contents is an array containing the values for all of the input columns.

Accuracy Evaluation

The Accuracy (%) is the ratio of correctly predicted observations and is formulated using the number of True positives (TP), True negatives (TN), False Positives (FP) and False negatives (FN),

$$Accuracy = 100 \times \frac{TP+TN}{TP+TN+FP+FN}$$

The accuracy of our model trained using Logistic Regression is 79%.

EXPERIMENTATION RESULTS

The algorithm classifies the comments based on polarity and provides insights as to what the type of video is, in the form of a uni-coded strings represented as emoticons.

No. of Videos	Reaction(Comments)	Reaction(likes/dislikes)
2037	😀	😀
133	😠	😀
37	😶	😀
35	😠	😠
17	😀	😠
3	😀	😶
1	😶	😶
2	😶	😠
0	😠	😶

Figure 9: Output of Analysis on polarity of videos

We have assigned Grinning Face emoticon for a positive reaction, Face Without Mouth emoticon for a neutral reaction, Angry Face emoticon for negative reaction. For most videos, the reaction with respect to comments and likes/dislikes is same, which is positive. But for 133 videos though based on likes it is a positive video, the comments state otherwise which is misleading. And for 17 videos, a greater number of dislikes state that the videos are negative but their comments state otherwise.

The outputs of SparkML Logistic Regression are as follows:

	video_id	polarity	prediction
0	vo391vKvD4s	1	1.0
1	k5qKGNeRb68	0	0.0
2	NFHAAtVkbpA	0	0.0
3	PpGyVCu2n9g	0	0.0
4	KMZ2vfloA9k	0	0.0
...	...	...	...
1807	33Z-ix1VLzI	0	0.0
1808	xEKSJLaiZa0	1	1.0
1809	q8b5mciIDnA	1	1.0
1810	v90yrcg6q9I	0	0.0
1811	6v3BWoddSgk	0	0.0

Figure 10: Output of prediction using LR

Also, performed tag analysis and from the packed bubble visualization it is evident that during the time data is collected, comedy and funny videos grab the most user views. Any YouTuber can choose video content with the help of trend visualization. We embedded this tableau visualization in jupyter notebook.

Figure 11: Output of Analysis on Tags of videos

CONCLUSION

The intention of this project is to make a preliminary examination on the relevance of comment text with the polarity of the video. An Analysis is also performed on two machine learning techniques, Textblob and Logistic regression, to classify the comments. Determined the top tags, to help YouTubers to choose content accordingly, for more views.

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] Bishop, C. (2006). Pattern Recognition and Machine Learning. Ney York: Spring- Verlag

[3] 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.

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

[5] Choudhury, Smitashree, and John G. Breslin. "User sentiment detection: a YouTube use case." (2010).

[6] Siersdorfer, S., Chelaru, S., Nejdl, W., & San Pedro, J. How useful are your comments?: analyzing and predicting youtube comments and comment ratings. ACM. In Proceedings of the 19th international conference on World wide web. pp. 891-900, (2010)

[7] Agrawal, S. Using syntactic and contextual information for sentiment polarity analysis. In Proceedings of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human. ACM. pp. 620-623. (2009).

[8] John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Do- main adaptation for sentiment classification. In Proceed- ings of ACL.

[9] Minqing Hu and Bing Liu. 2004. Mining opinion features in customer reviews. In Proceedings of the 19th national conference on Artifical intelligence, AAAI'04, pages 755–760. AAAI Press.

[10] Raj, A. (2018). Sentiment Analysis of the Nostalgic Comments on the songs of 20th Century from YouTube. Maynooth University, Computer Science. Maynooth: Unpublished MSc thesis.

[11] Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 5(4), 1093-1113.

[12] Mitchell J., https://www.kaggle.com/datasnaek/youtube

[13] G. Mohana Prabha, B. Madhumitha, R. P. Ramya, April 2019, Predicting the Popularity of Trending Videos in Youtube Using Sentimental Analysis, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8, Issue-6S3, April 2019