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 source 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, and rating (like/views) 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. To minimize this issue, we present a sentimental analysis approach to user comments using Natural Language Processing. This project offers a novel approach to analyze opinions posted by users about a particular video.

Keywords

YouTube; Sentiment; Analysis; Comments; Mining; Views; Likes; Dislikes; 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. This 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 particular 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 a 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.

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).

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 machine learning could be applied to determine the proportion of relevant comments. The added value from this could facilitate further sociological investigations into a modern phenomenon and be useful for a proper search.

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. 2

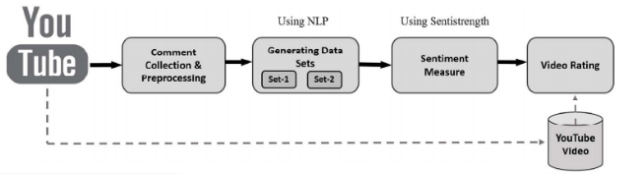

Figure 1: Overall work process of sentiment analysis on user comments

First, comment collection is done from the dataset repository and pre-processing module extracts data (comments) from the specific YouTube video and do some language pre-processing to prepare for the next process. Second, 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, apply the Standard Deviation to get the rating result and visualization is performed. In detail the methodologies are explained below.

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

The YouTube API is not effective at formatting comments by relevance, although it claims to do so. As a result, the most relevant comments do not align with the top comments at all, they aren't even sorted by likes or replies.

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

Preprosessing

Remove all the expressions which are irrelevant for the proposed methodology like date ("Dec 2-2010"or "2-12-2010"), link (www.abc.com, www.pqr.com etc.), numbers (192, 500 etc.) and special characters ("*","/","!","","?","#","&","$"), emoticons (":), :D, <3, ;) etc.") and different language (Chinese, Cantonese, Telugu , Hindi etc.). Remove all the punctuation marks such as period (“.”), space ("-"), commas (","), semicolon (“;”), hash (“-“) etc.

K-Fold Cross Validation: Cross-validation is the best way to stretch the validity of the manually annotated data since it enables it to strain on a large number of the documents. It is perfect for checking model effectiveness, especially when there is a need to mitigate for overfitting. In k-fold cross-validation, the data is divided into k different subsets. One of the k subsets is used as the test data, and the remaining data k-1 are taken as the training data. The overall error obtained as the average of all the k trials and is a measure of the total effectiveness of the model. Every data point gets to be in a validation set exactly once and gets to be in a training set k-1 times. This significantly reduces bias as most of the data is used for fitting and lowers the variance, as most of the data is also being used in the validation set.

NLP Algorithm for generating subsets

For each evaluating video, two datasets are made according to the proposed method. Both are made from the processed comments text. To make datasets first, in the processed text, remove all the stop words and then convert all the words into their singular form and thus make dataset 1. Next, for Dataset 2, all the adjectives (essential words of the comment text) of the comments are gathered. Empirically and from the self-analysis, on YouTube video comments, it seems that adjectives are the most critical indicators for a user's feeling and decision about the video's quality and relevancy.

Machine Learning Algorithms

Naïve Bayes classifier: Naïve Bayes is a classifier that uses Bayes theorem. Bayes theorem. The implementation of the Naïve Bayes classifier assumes that the data instances are conditionally independent in order to compute the MAP hypothesis. It is a well-known, straightforward algorithm and is not as computationally demanding as other approaches.

Decision Tree Classifier: Decision trees are a widely used algorithm in machine learning since they can be adapted easily to any data. The algorithm is mainly used when there is a need for many hierarchical distinctions. The tree itself is represented as a linear structure and can be easily understood. The decision tree consists of a root node, representing the entire data and decision nodes and leaf nodes, which illustrate the classification. The data is passed through the tree to classify the instance. At each of the decision nodes, a particular feature from the input is compared with a constant that is recognized during the training phase. The data will pass through all these decision nodes until it reaches a leaf node that represents the particular assigned class.

Random Forest Classifier: Random forest classifier creates a set of decision trees from a randomly selected subset of

the training set. It then clusters the votes from different decision trees to decide the final class of the test object. The fundamental concept behind random forest is a simple but powerful one. The low correlation between models is the key. Just like how investments with low correlations (like stocks and bonds) come together to form a container that is greater than the sum of its parts, uncorrelated models can produce ensemble predictions that are more accurate than any of the individual predictions. While some trees may be wrong, many other trees will be right, so as a group, the trees can move in the correct direction.

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),

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

VISUALIZATIONS

Tableau is used for visualizing the end results and the estimated visualization is as follows:

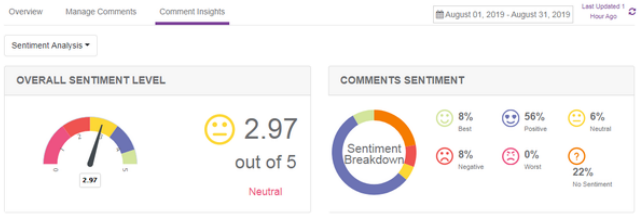

Image Source https://www.videoamigo.com/youtube-creators/youtube-comment-manager.

CONCLUSION

This intention of this project is to make a preliminary examination of the application of Machine learning techniques to comments gathered from videos on YouTube to determine how well they could identify relevant videos based on polarity of comments. Three machine learning algorithms will be used: Naïve Bayes, Decision Tree and Random Forest. Additionally, K-fold cross- validation will be used along with NLP algorithm for generating subsets.The comparison in accuracy of different machine learning techniques will be established using visualizations.

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