Proc. of the 2017 IEEE International Conference on Signal and Image Processing Applications (IEEE ICSIPA 2017), Malaysia, September 12-14, 2017

Retrieving YouTube Video by Sentiment Analysis on User Comment

Hanif Bhuiyan
Department of CSE
University of Asia Pacific
74/A, Green Road, Dhaka
hanif_tushar.cse@uap-bd.edu

Jinat Ara
Department of CSE
Southeast University
Banani, Dhaka
aracse2014@gmail.com

Rajon Bardhan
Department of CSE
Southeast University
Banani, Dhaka
rajonbardhan@gmail.com

Md. Rashedul Islam
Department of CSE
University of Asia Pacific
74/A, Green Road, Dhaka
rashed.cse@uap-bd.edu

Abstract— **YouTube is one of the comprehensive video information source on the web where video is uploading continuously in real time. It is one of the most popular site in social media, where users interact with sharing, commenting and rating (like/views) videos. Generally the quality, relevancy 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 Natural Language processing (NLP) based sentiment analysis approach on user comments. This analysis helps to find out the most relevant and popular video of YouTube according to the search. The effectiveness of the proposed scheme has been proved by a data driven experiment in terms of accuracy of finding relevant, popular and high quality video.**

Keywords- YouTube; Comment; Sentiment analysis

I. INTRODUCTION

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 rapidly increasing popularity, these sites have become a source of massive amount of real time data of videos, images etc. Among them, YouTube[1] is one of the world's largest video sharing platforms, where videos are uploading continuously by the millions of users (companies, private persons, etc.) [1]. YouTube has emerged as a comprehensive and accessible compilation of video information source on the web. It is a unique environment with many facets such as multi-modal, multi-lingual, multi-domain and multi-cultural [1]. This versatility of variety and attractive shared content draw the 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 on Dec, 2016 [2]. In order to increase the user's interaction it allows their users to express their opinion by rating the viewed objects (via clicking on the like/dislike buttons) and interacting with the other community members (via the comments feature) [3]. These activities (like / dislike /number of views) of the users can serve as a global indicator of quality or popularity for a particular video [3, 4]. Moreover, these Meta data (like/dislike/number of views) serve the purpose of helping the community to filter relevant opinions more efficiently [3, 5, 6]. When we search for a specific video through some keyword on specific 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 an example when we do a query like "The Amazing Spider-Man 2". For that search, Figure. 1 represents the result where none of the videos actually that movie. Most of the result for this query are inconsistent. Among the result some are movie's cut scenes, 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, in order 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). In order to solve these issues several works have done where some groups have been reported significant progress [3, 4, 7, 8].

The research community continuously showed keen interest in analyzing and exploiting the rich content shared on YouTube. Some earlier studies attempted to investigate the retrieval potential of the video using not only the Meta data (like/dislike/number of views) [1] but also using comments [3, 7]. In addition, from the self-observation it comes out that in YouTube, an appropriate video carry a high amount of positive comments rather than an inappropriate video. Therefore, it is speculated that comments might one of the vital source to perceive about the video quality, perfection, relevancy and its popularity. However, user comments usually found in unstructured way and it is quite difficult to analyze. Therefore, in this paper we present a Natural Language Processing (NLP) based sentiment analysis approach in order to find the popular video which is relevant. In this study, the answer is seek that, how useful is this metadata (comments), to improve the retrieval effectiveness?

[1] https://www.youtube.com

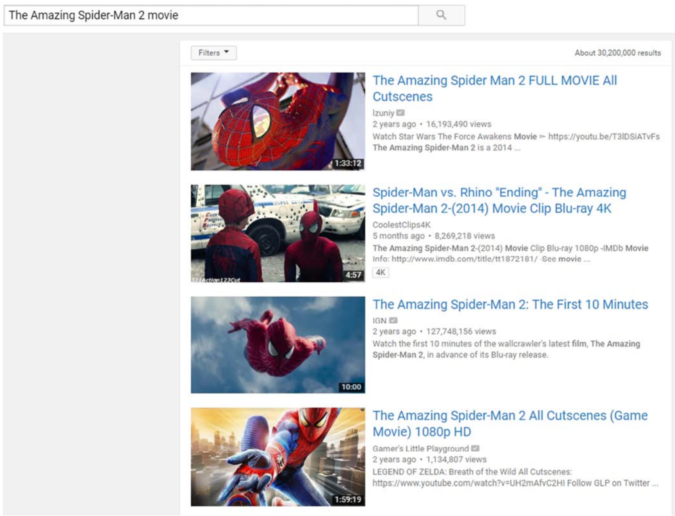

Fig. 1. Searching in YouTube

Moreover, the experimental result shows that metadata (comments) have the potential to correctly mention the relevancy or popularity of the video. The brief description of this proposed process is stated in the section III.

II. Related Work

Several research has been undertaken on different aspects of YouTube video features [5]. Among them comments are one of the important one to make a decision (comment rating, topic categories etc.,) about the particular video [6]. These comments are also used to annotate the video object [7, 9]. Comments also reflect the user's behavior and could use to find the troll users [1]. Moreover, by analyzing the sentiment of comments it is possible to find the users positivity or negativity about video [10]. Based on comments, researchers categorize the videos in several category [11, 12]. Furthermore, for improving video retrieval process Altingovde et al. proposed a method based on basic feature and social feature [3]. Lehner et al. Work on YouTube video comments, like, dislikes for showing that user's perception (like/dislike) are influenced by valuable comments [4]. These two methods worked to find the popularity of video using various features so that it could help to retrieve the useful video. Although these two proposed approaches [3, 4] showed impressive work for video retrieval process but they used like/dislike and views. Sometimes which may lead to inaccurate result. On the contrary, we only analyze a large amount of comments instead of others features (like/views etc.) for finding relevant video which might be useful for YouTube users.

III. Methodology

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. First, comment collection and preprocessing module extracts data (comments) from the specific YouTube video and do some language preprocessing to prepare for the next process. Second, the processed text go through a NLP-based methods to generate data sets. Next, apply the sentiment classifier (Sentistrength) on the data sets to calculate the positivity and negativity scores. Finally, apply the Standard Deviation to get the rating result. In detail the methodology are explained in below.

A. Comment Collection and Preprocessing:

The goal of this section is acquisition of comment of a selected YouTube video. In order to address this task a focused crawler is implemented. According to the video URL, it extracts 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 preprocessing on these unstructured comments to generate the data sets.

After extracting the comments, following changes are performed:

- Remove all the expressions which are irrelevant for the proposed methodology like date ("Dec 2-2010"or "2-12-2010"), link (www.imdb.com, www.tmdb.com etc.), numbers (12, 20 etc.) and special characters ("*","/","!","@","?","#","&","$"), emoticons ("☺, ☹, <3 etc.") and different language (Chinese, Arabic, Bangla , Hindi etc.).

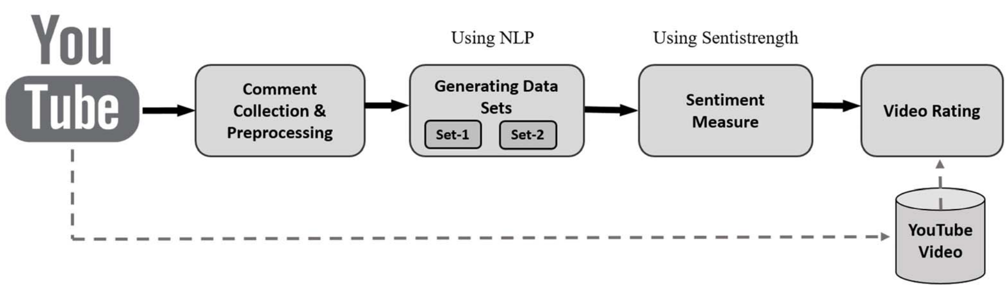

Fig. 2. Overall work process of sentiment analysis on user comment.

- Remove all the punctuations such as period ("."), space ("-"), commas (","), semicolon (";"), hash ("-") etc.

B. Generating Data Sets:

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 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 1. Next, for Dataset 2 all the adjectives [14] (important 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 important indicators for a user's feeling and decision about the video's quality and relevancy. Therefore, to make the list of adjectives of the comments Stanford Part-of-Speech tagger (POS Tagger)[2] is applied and identify all the adjective words and thus make dataset 2.

C. Sentiment Measure:

In this section, we use SentiStrength[3] thesaurus in both the dataset to figure out the overall sentiment of user comments. SentiStrength is a sentiment lexicon analysis classifier which estimates the strength of positive and negative sentiment of the comment words. It reports two sentiment strengths:- 1 (not negative) to -5 (extremely negative) and 1 (not positive) to 5 (extremely 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. For an example, Table I shows the sentiment measure for a particular comment ("He changed the world as we know it and still remains super humble. Much respect.") according to the proposed method. From Table I, it seems that, what that comment is meaning about the video, is it positive or negative. In overall, after calculating the sentiment value of all comments the standard deviation value is calculated (by applying standard deviation technique, describe in next section) of these values for next step of the process.

Table I: Sentiment measure of a comment

Word	pos	neg
changed	1	-1
world	1	-1
remains	1	-1
super	1	-1
humble	1	-1
respect	3	-1

D. Video Rating:

In this section the statistical comparison of the standard deviation (SD) values is conducted (both positive and negative distributions) across the both dataset. After measuring the sentiment value of both data set the standard deviation technique is applied. The analysis of SD value for each group of data systematically resulted in a strong support of the significance of the difference across the two groups. For an example, here Figure. 3(a) and Figure. 3(b) represent the scenario of determining SD value for 10 videos for both types of dataset. These graphs depicts the SD values for negativity and positivity which revealing that negative sentiment values are predominant in negatively rated comments, whereas positive sentiment values are predominant in positively rated comments. For each video, there are two SD value come from two data set. Then, the calculated average SD value determined the relevancy, perfection and quality of the video.

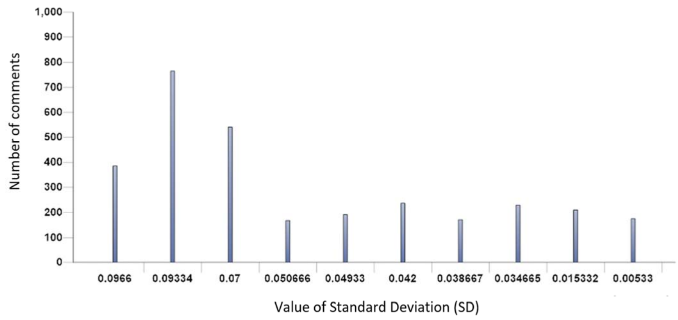

Fig. 3(a). SD value for dataset 1

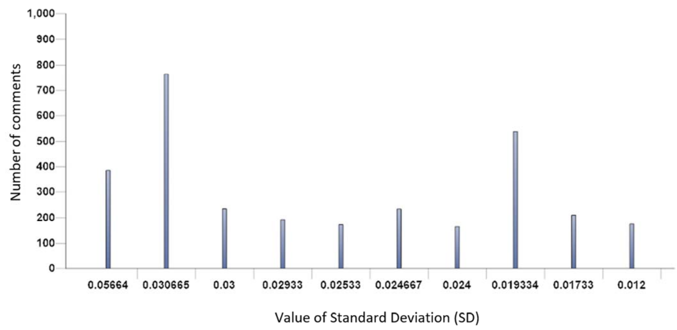

Fig. 3(b). SD value for dataset 2

IV. EXPERIMENT

This section presents the experiment results of the proposed sentiment analysis approach. To evaluate the proposed approach the experiment is conducted on 1000 videos of YouTube which were selected randomly. However, precisely 10 categories (education, science and technology, entertainment, cartoon, etc.) were selected for those videos. For each category 100 videos were considered and for each video, 1,000 comments were considered.

Before experiment, a manual inspection was performed on randomly selected 100 videos of YouTube. 10 volunteers were recruited to justify the result of the video through the proposed process. Volunteers checked the relevancy and quality of the video. Result suggested that individual video dataset having > 0.5 average SD is perfect and relevant

[2] http://nlp.stanford.edu/software/tagger.shtml
[3] http://sentistrength.wlv.ac.uk/

Video category	Number of videos	Accuracy when considering each dataset		Accuracy when considering both dataset together
Education	100	Dataset1	73.66%	62.125%
		Dataset2	50.59%	
Science and technology	100	Dataset1	86.88%	75.435%
		Dataset2	63.99%	
Entertainment	100	Dataset1	82.33%	68.945%
		Dataset2	55.56%	
Pets and animals	100	Dataset1	83.77%	66.38%
		Dataset2	48.99%	
Cartoon	100	Dataset1	64.55%	51.21%
		Dataset2	37.87%	
Documentary	100	Dataset1	79.56%	64.665%
		Dataset2	49.77%	
Movie	100	Dataset1	77.24%	63.615%
		Dataset2	47.99%	
News and politics	100	Dataset1	80.24%	57.565%
		Dataset2	34.89%	
Games and animation	100	Dataset1	69.85%	57.755%
		Dataset2	45.66%	
Song	100	Dataset1	76.54%	61.216%
		Dataset2	45.883%	

Table II: Accuracy of finding relevant video based on comments

according to the search. It was also found that the videos are of high quality, perfect, relevant and popular when the average SD become > 0.5 considering both dataset together. Therefore for the experiment of the proposed method the threshold value was set 0.5. Experiment result is shown in Table II.

The result shows that in order to find the relevant and popular video in YouTube using comments are very effective. From the Table II, it seems, for dataset 1 the result is much better than for dataset 2. The highest accuracy 75.435% of the proposed approach was occurred for science and technology related videos where lowest accuracy 51.21% has been calculated for cartoon categories of video. In addition, for dataset 1 the average accuracy of the proposed approach was 77.462% when whole comments were considered. Although the accuracy was low for dataset 2 but using adjectives gives much more accurate result sometimes. Therefore, considering of both dataset could give the real picture of a particular video than using specific one dataset. For an example in Table III, when a video was searched like “Huawei Ascend p7 Review ", in the search panel it was shown that first few videos were relevant and perfect. But to determine which one was the best and more popular and

Table III: An example of measuring video relevancy and quality through the proposed method

Video	Dataset	Accuracy when considering each dataset	Accuracy when considering both dataset together
Video1	Dataset1	1.134411513	0.866642
	Dataset2	0.598874	
Video2	Dataset1	1.120566227	0.939081
	Dataset2	0.757596	

which one should be appear first. For this, analyzing the both dataset was essential. From the result in Table III, it seems that first video of dataset 1 gives better accuracy than second video of dataset 1. However, accuracy of dataset 2 of second video is much better than dataset 2 of first video. Therefore, in such case considering both dataset gives us the real picture of those videos quality and perfection. Like, in this case second video is the perfect and much relevant than first one. So if both datasets is considered to find the video on YouTube then the result might be much better.

From the result it can be concluded that, if the analysis of YouTube video performed based on the comments text, then our proposed approach might commit good result. However, it depends as much as semantically the text is analyzed, the more we analyze the text the more chance to get better accuracy.

V. CONCLUSION

This paper illustrates an automatic process for finding useful video by sentiment analysis of user’s comments based on Natural Language Processing (NLP). Our approach evaluated the quality, relevancy and popularity of YouTube videos considering the relationship of user’s sentiments expressed in comments. We analyzed a sample of almost 1 million YouTube comments. Large-scale studies of YouTube video Meta data (comment) using the NLP and SentiStrength revealed the importance of user sentiments. The experimental result shows the efficiency of the proposed approach by revealing maximum 75.435% accuracy in order to retrieving effective video and also provides the direction to pursue the work to do more analysis on comment.

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] http:// www.wikipedia .com.

[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] 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.

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

[7] E. Momeni, C. Cardie and M. Ott, "Properties, Prediction, and Prevalence of Useful User-Generated Comments for Descriptive Annotation of Social Media Objects," In Proceedings of ICWSM, 2013.

[8] O. Uryupina, B. Plank, A. Severyn, A. Rotondi and A. Moschitti, "SenTube: A Corpus for Sentiment Analysis on YouTube Social Media," In LREC, 2014, pp. 4244-4249.

[9] E. Momeni, B. Haslhofer, K. Tao and G. J. Houben, "Sifting useful comments from Flickr Commons and YouTube," International Journal on Digital Libraries, 16(2), 2015, pp.161-179.

[10] H. Lee, Y. Han, Y. Kim and K. Kim, "Sentiment analysis on online social network using probability Model," In Proceedings of the Sixth International Conference on Advances in Future Internet, 2014, pp. 14-19.

[11] K. Filippova and K. B. Hall, "Improved video categorization from text metadata and user comments," In Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, 2011, pp. 835-842.

[12] A. Severyn, A. Moschitti, O. Uryupina, B. Plank and K. Filippova, "Opinion Mining on YouTube," In ACL (1), 2014, pp. 1252-1261.

[13] Z. Wu and E. Ito, "Correlation analysis between user's emotional comments and popularity measures," In Advanced Applied Informatics (IIAIAAI), IIAI 3rd International Conference (IEEE), 2014, pp. 280-283.

[14] H. Bhuiyan, K.J. Oh, M.k. Hong and G.S. Jo "An unsupervised approach for identifying the Infobox template of wikipedia article." In 18th International Conference on Computational Science and Engineering (CSE), 2015 IEEE, pp. 334-338.