Exploring The User Comments From Youtube Videos Using NLP and ML

[1]Kudupudi Vijaya Durga, [2]D.S.Ramkiran
[1,2]Dept. of Computer Science & Engineering, KIET, Kakinada, AP, India

Abstract
Sentiment analysis is a process that discovers the user opinions and views against any service or a product. YouTube is one of the most popular videos sharing platforms obtaining millions of views. These receive several comments, containing valuable information that helps in improving the rating levels of the uploaded content. These comments are utilized by using natural language processing techniques and machine learning techniques. There are many attempts had been proposed scholarly with two (positive or negative), three (two with neutral) or multiple (happy, sad, fear, surprise and anger) classes. However, it is challenging to choose the best accurate model. Therefore, there had been attempts to use sentiment analysis on YouTube comments in identifying the polarity as well. This research paper investigates the sentiment analysis methods and techniques that can be used on the YouTube content. Additionally, it explains and categorizes these approaches which are useful in researches in data mining and sentiment analysis.

I. Introduction

Google Video content grew so popular that Google acquired YouTube in 2006, just one year after YouTube was founded. Today, over 300 hours of video are uploaded to YouTube every minute and almost 5 billion videos are being watched each day. YouTube introduced the slogan "Broadcast Yourself" as a way to encourage the everyday person to put their life on film and host it on YouTube. Since the first video posted by one of YouTube's founder, titled "Me at the Zoo," billions of hours have been shared about makeup, gaming, technology trends, and everything in between. Monetizing content was implemented in 2007, and brands have taken advantage of the site's profitability by partnering with popular influencers to access their audiences. The emergence of user-generated content, specifically through videos on YouTube, has drastically changed the way marketers advertise to their audiences. Today, YouTube's landscape could be better explained as "Broadcast Brands," as companies and creators are tapping into the world of YouTube advertising and sponsored content. Little research has been published in understanding the implications of brandsponsored content and brand strategy specifically through YouTube. With such information, brands will more effectively be able to leverage YouTube as a way to authentically engage with users to foster a symbiotic relationship between the brand, the content creator, and the audience. The primary goal of this thesis was to gain a well-rounded understanding of YouTube and how it is viewed in the marketplace. This thesis seeks to delve into 7 literature surrounding YouTube's history, current landscape, and competitors in order to understand where it stands in the market. This thesis also seeks to pair this understanding with results derived from primary research. The primary research conducted draws conclusions from YouTube's two main user sets – businesses and millennial. By pairing key points derived from secondary literature with primary research trends about how users interact with YouTube, advertisers and business professionals can better understand how to leverage the platform. As well, YouTube, and ultimately Google, can tap into the findings about their user sets to understand how the market is viewing the platform, leverage its strengths, and develop tactics to address challenges the platform currently faces

II. Quality Of Youtube Comments

They write about the poor public image that YouTube comments has in social media, and that the users attach little or no value to the comments of a video. But the aim of the study was to use a comment classification approach that
captures the salient aspects of YouTube comments. P. Schultes et al. found that their classifiers is able to perform very fast lightweight semantic video analysis. And in addition to that, they find that a videos likes and dislikes are influenced by the distribution of valuable and invaluable comments.
The report How useful are your comments?: Analyzing and Predicting YouTube Comments and Comment Rating studies the correlation between comment sentiment and comment rating eg. like or dislike on the comment itself. Their study concluded that it is indeed possible to create a classifier that accurately predicts which comments are useful and which ones are not. Comments which are not deemed useful are ones that contain discriminatory language.

III. Sentiment Analysis On Youtube Comments

No earlier work on predicting a like/dislike ratio on YouTube videos based on comments was found while researching this topic. Though, research on YouTube comments has been published. Atte Oksanen et al. published a research paper on pro-anorexia and anti-anorexia videos on YouTube in 2015. The aim of their study was study emo-tional reactions to these anorexia-topic videos using sentiment anal-ysis. They analyzed the sentiments on comments in both pro- and anti-anorexia videos using ordinary least squares regression models. The results from the sentiment analysis show that anti-anorexia videos had both more positive comments and more likes than pro-anorexia videos. Similar to our work they count the comments with positive sentiment and base their conclusion on that. But unlike our comparison they are not trying to find a correlation between two sets of data based on their found emotional sentiment.

III. Data Collection Process And Algorithm

We modelled the data by automating queries and keyword based searches to gather videos and their corresponding comments. Python scripts using the YouTube APIs were used to extract information about each video (comments and their timestamps). We collected 1000 comments per video (YouTube allows a maximum of 1000 comments per video to be accessed through the APIs), and used keywords like "Federer", "Nadal", "Obama" etc., to collect the data for specific keywords. The timestamp and author name of each video were also collected. The final dataset used for the sentiment analysis had more than 3000 videos and more than 7 million comments. We performed data pre-processing on

the collected comments. YouTube comments comprise of several languages depending on the demography of the commenter. However, to simplify the sentiment analysis, we modified the data collection scripts to collect only English comments. From the collected English comments, only comments in the standard UTF-8 encoding were selected in order to remove comments with unwanted characters. The steps below explain the procedure to collect the comments with their respective timestamps and author names for the keywords specified by the user.

In steps 1-4, the Google APIs for YouTube are used to configure the query with the number of videos to be fetched, the language of interest for comments, the search keyword, and how the comments are to be sorted.
Step 5 collects the IDs of the videos related to the specified keyword.
Steps 6 and 7 collect the comments associated with these videos and extract the timestamps, author names and comment text from the comment entries. All the comments for a single keyword are aggregated into one dataset which is used as the test set as explained in the following:
Step 1: Prompt the user to specify the search keyword (keywords) and number of videos (numVideos)
Step 2: Set maxNumVideos = max(50; num Videos) (As Google limits the maximum number of videos fetched in one iteration to 50)
Step 3: Set up the YouTube client to use the YouTube-Service() API to communicate with the YouTube servers Step 4: Use the YouTubeVideoQuery() API to set the query parameters like language, search keyword,etc
Step 5: Perform successive queries to get the video ID of each video related to the keyword Step 6: Collect the comments associated with each video ID using the Get YouTube Video Comment Feed() API (maximum limit of comments per video is 1000)
Step 7: Extract the comments with their respective timestamps and author names an input into one of two or more discrete classes. An example of a classification problem with two classes is determining whether an image contains a cat or not. Classifiers is a learning algorithm combined with a set of training data [1]. The classifier is trained using the
training data which consists of al-ready classified inputs. For example a set of pictures which are labelled with "contains cat" or
"doesn't contain cat".

IV. Classifying Using Machine Learning

A common way of creating a classifier for text is by using a machine learning algorithm. The algorithm will use the training data to find patterns between the given inputs and their sentiment. The problem with using text as an input for machine learning algorithms is that it doesn't work. Text has to be converted into numbers or vectors of some kind. This conversion is called feature extraction [7].

Table 1: Vector representation of a sentence

cats	are	Very	cute	I	think	about
1	0	0	1	1	1	1

Which is the vector (1, 0, 0, 1, 1, 1, 1). However, most practical applications of the BoW model will give us vectors with vastly larger dimensions since it needs to contain all words known to the model. Thus, a vector representation of our sentence, "I think about cute cats", given a corpus of all the words in the Oxford Dictionary would have 171,476 [8] dimensions where all except 5 values are "0".
Decision rules can be generated by constructing association rules with the target variable on the right. They can also denote temporal or causal relations. Decision tree using flowchart symbols Commonly a decision tree is drawn using flowchart symbols as it is easier for many to read and understand.

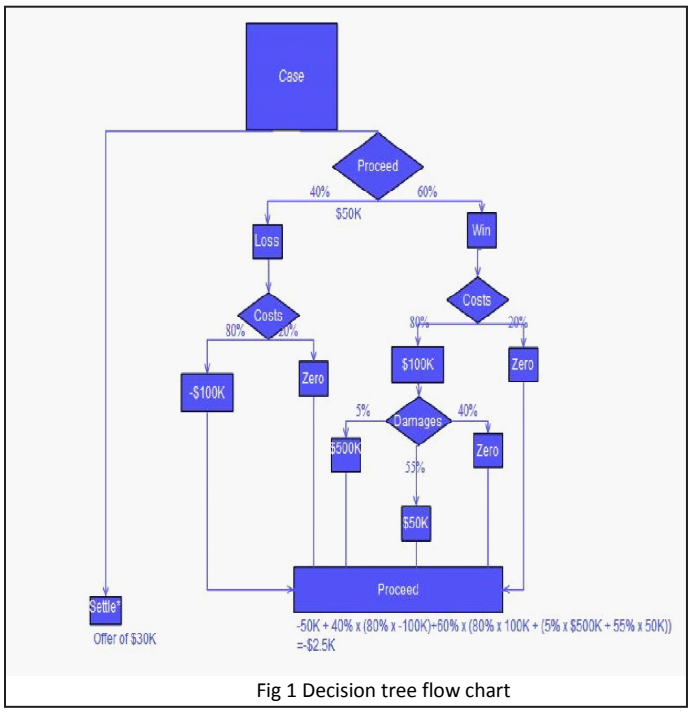


Fig 1 Decision tree flow chart

Analysis example Analysis can take into account the decision maker's (e.g., the company's) preference or utility function, for example:

V. Implementation And Evaluation

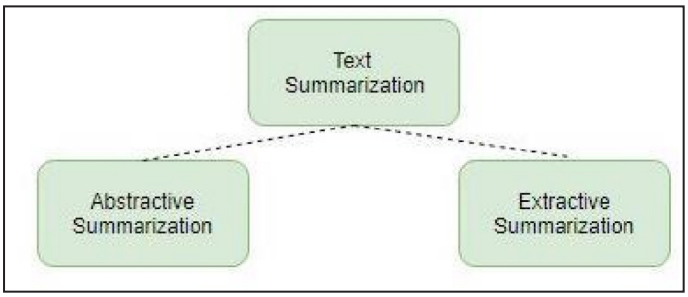
To implement NB and knn we use Weka [31], a data mining suite that implements a wide variety of machine learning and classification techniques

```
collection <include.DocumentCollection.DocumentCollection instance at 0x11bbe5998>
Results for -- K Nearest Neighbours -- classifier over  10 Folds -  Direct values
Confusion Matrix Cluster: See Report for Cluster (too large for terminal output)
Avg K-Fold Classification Error Rate:  33.84766287595271 %
Avg F1 Score:  77.75594365250515 %
Avg Precision:  66.39359868685489 %
Avg Recall:  95.34808278813992 %

Results for -- K Nearest Neighbours -- classifier over  10 Folds -  1-gram and 2-gram
Confusion Matrix Cluster: See Report for Cluster (too large for terminal output)
Avg K-Fold Classification Error Rate:  31.120738988016182 %
Avg F1 Score:  80.96809569549607 %
Avg Precision:  69.18773008169656 %
Avg Recall:  98.79810017834261 %

Results for -- K Nearest Neighbours -- classifier over  10 Folds -  1-gram and 2-gram & TF-IDF
Confusion Matrix Cluster: See Report for Cluster (too large for terminal output)
Avg K-Fold Classification Error Rate:  27.411008523070223 %
Avg F1 Score:  82.39612107377556 %
Avg Precision:  74.46944078555474 %
Avg Recall:  93.1979229033959 %

collection <include.DocumentCollection.DocumentCollection instance at 0x11bbe5998>
Results for -- MultiNomial Naive Bayes -- classifier over  10 Folds -  Direct values
Confusion Matrix Cluster: See Report for Cluster (too large for terminal output)
Avg K-Fold Classification Error Rate:  20.891156127516087 %
Avg F1 Score:  85.93815331119158 %
Avg Precision:  81.33158366225366 %
Avg Recall:  91.90019998574769 %

Results for -- MultiNomial Naive Bayes -- classifier over  10 Folds -  1-gram and 2-gram
Confusion Matrix Cluster: See Report for Cluster (too large for terminal output)
Avg K-Fold Classification Error Rate:  25.25206300150502 %
Avg F1 Score:  82.14303967374002 %
Avg Precision:  75.96563399657283 %
Avg Recall:  90.72543217062517 %

Results for -- MultiNomial Naive Bayes -- classifier over  10 Folds -  1-gram and 2-gram & TF-IDF
Confusion Matrix Cluster: See Report for Cluster (too large for terminal output)
Avg K-Fold Classification Error Rate:  28.178894173889958 %
Avg F1 Score:  82.32957519592881 %
Avg Precision:  71.68835352905243 %
Avg Recall:  98.11004272156316 %
```

VI. Comments Summarization

Text Summarization refers to the shortening technique of long pieces of text. The aim is to create a coherent and smooth summary with only the main points outlined in the document. Automatic text summary is a common problem in the processing of machine learning and natural language.

There are generally two types of Summarization, abstract and extractive Summarization.



Abstractive Summarization: Abstract methods select words based on semanthropy, even those words did not appear in the source documents. In a new way, it aims at producing substantial material. Using advanced natural language techniques, they interpret and examine the text to create a new short text conveying the most critical information from the original text.

VII. Conclusion

There is some correlation between the percentage of likes and percent-age of positive comments on YouTube. Though, since the variation is very high it is not possible using the comment sentiment alone to ac-curately predict the percentage of likes using our method. The answer to the research question, Can the comments on a YouTube video be used to determine what ratio of the viewers liked or disliked the video using senti-ment analysis? is difficult to answer using only our results. The method needs to improve in order to draw any substantial conclusions. Most importantly does the training data need to improve, irrelevant com-ments be sorted out and the amount of videos to analyze increase. If these areas are improved upon, a conclusive answer to our research question could be found.

VIII. Further Research

Our future work will include conducting surveys and human experiments with actual You Tubers to get a more in-depth understanding of their needs and expectations as well as response patterns to their viewers If we were to continue this research and improve upon it, we could take the following points into consideration. Improving the training data. The training data used in this report is one of the big flaws. The main issue of the training data was that it was a twitter training data set. Even though tweets are microblog posts and similar to YouTube comments in length, the language is very different. The difference in language can cause the model to wrongly classify comments. Also, the comments have been automatically generated which mean that they are not guaranteed to be completely accurate.

References

[1] T. C. Alberto, V. L. Johannes and A. A. Tiago, "Tubespam: Comment spam filtering on youtube", IEEE 14th International Conference on Machine Learning and Applications (ICMLA) Miami, FL, USA, pp. 138-143, 2015.

[2] Hammad Afzal, Robert Stevens, and Goran Nenadic, "Towards semantic annotation of bioinformatics services: building a controlled vocabulary", Third International Symposium on Semantic Mining in Biomedicine, pp. 5-12, 2008.

[3] Tiago A. Almeidaa, Tiago P. Silvaa, Igor Santosb, Jos´e M. G´omez Hidalgoca, "Text Normalization and Semantic Indexing to Enhance Instant Messaging and SMS Spam Filtering", Knowledge-Based Systems, Vol. 108, pp. 25-32, 2016.

[4] Zakia Zaman, Sadia Sharmin, "Spam Detection in Social Media Employing Machine Learning Tool for Text Mining" 13tn International Conference On Signal Image Technology & internet based system (SITIS) 2017.

[5] Elizabeth Poché , Nishant Jha, Grant Williams, Jazmine Staten, Miles Vesper, Anas Mahmoud " Analyzing User Comments on YouTube Coding Tutorial Videos", IEEE 25th International Conference on Program Comprehension (ICPC), 2017.

[6] Shreyas Aiyara, Nisha P Shettyb "N-Gram Assisted Youtube Spam Comment Detection" International Conference on Computational Intelligence and Data Science (ICCIDS), 2018

[7] Arif Mehmood, Byung-Won On, Ingyu Lee, Imran Ashraf, Gyu Sang Choi "Spam comments prediction using stacking with ensemble learning", 10th International Conference on Computer and Electrical Engineering, 2018.

[8] Alper Kursat Uysal "Feature Selection for Comment Spam Filtering on YouTube", 10th International Conference on Computer and Electrical Engineering, 2018.

[9] A. O. Abdullah, M. A. Ali, M. Karabatak, and A. Sengur, "A comparative analysis of common YouTube comment spam filtering techniques", Digital Forensic and Security (ISDFS), 6th International Symposium on, 2018, pp. 1-5: IEEE.

[10] M. Carlisle, "Using YouTube to enhance student class preparation in an introductory java course," Proceedings of the 41st ACM Technical Symposium on Computer Science Education, pp. 470–474, 2010.

[11] C. Meda, F. Bisio, P. Gastaldo, and R. Zunino, "A machine learning approach for Twitter spammers detection," International Carnahan Conference on Security Technology (ICCST), pp. 13-16, Oct. 2014. 12.C. Alberto, Tulio and Lochter, Johannes and Almeida, Tiago (2015). "TubeSpam: Comment Spam Filtering on YouTube." 138-143.

[12] 10.1109/ICMLA.2015.37.

[13] Stefan Siersdorfer and Sergiu Chelaru, "How useful are your comments?: analyzing and predicting youtube comments and comment ratings". In Proceedings of the 19th international conference on World wide web, 2010, pp. 891-900.

[14] A. Ammari, et al., "Identifying relevant youtube comments to derive socially augmented user models: a semantically enriched machine learning approach," Book Identifying relevant youtube comments to derive socially augmented user models: a semantically enriched machine learning approach, Series Identifying relevant youtube comments to derive socially augmented user models: a semantically enriched machine learning approach, ed., Editor ed.^eds., Springer-Verlag, 2012, pp. 71- 85.

[15] R. Chowdury, M. Monsur Adnan, G. Mahmud, and R. Rahman, "A data mining based spam detection system for youtube," in Digital Informa- tion Management (ICDIM), 2013 Eighth International Conference on, Sept 2013, pp. 373–378.

[16] Serbanoiu, A., Rebedea T., "Relevance-Based Ranking of Video Comments on YouTube". In CSCS '13 Proceedings of the 2013 19th International Conference on Control Systems and Computer Science, 2013, Washington, USA, pp.225-231.

[17] C. Rădulescu, M. Dinsoreanu, and R. Potolea, "Identification of spam comments using natural language processing techniques," in Intelligent Computer Communication and Processing (ICCP), 2014 IEEE International Conference on, 2014, pp. 29-35: IEEE.