

IA5050
Data Mining in Cyberspace
Spring 2015

YouTube Reactions Gauged by Comments

Alejandro Mera
Mohanaivel Senthilvel
Nicolas Gengo
Tawa Gonzalez

Table Of Contents

	Page No.
I. Business Problem.....	2
II. Data Sources.....	3
III. Data Collection.....	3
IV. Data Munging.....	6
V. Analysis.....	7
VI. Results.....	12
VII. Risk Assessment.....	17
VIII. Conclusion.....	19
IX. Appendix.....	21

Business Problem

Online videos have always been utilized by companies for gathering potential customers. These videos, however, have resulted in either positive or negative feedback. With the advent of modern technology and the Internet, companies have been exposed to and adapted innovative ways to achieve more positive results from their online video campaigns. Social networking has been a major contributor to product purchases, as people share products which satisfy them. Popular products gather positive feedback. Therefore, companies that desire higher product sales benefit from generating positive results from their videos. Currently, explosively popular videos are referred to as viral videos.

Breaking down the data behind online video campaigns enables companies to tease out the emotional impact of a video providing insights as to how the customers are receiving their product videos. However, the lack of “in-house” analytic expertise presents a challenge in this area. With the advent of big data analytics, companies are realizing that data mining contributes towards filling this gap. As an analytics service company, our mission is to help clients refine their online video marketing campaigns. In the market research industry, our service is used as a focus group tool to help our clients understand how their online videos are being received by the public; how to make improvements in those videos; how to reach a larger audience; and, ultimately, how to achieve viral status.

The purpose of this project is to create an analysis that will identify the range of positive and negative reactions that videos provoke, and provide our services, recommendations, and suggestions for achieving viral status.

Data Sources

Popular YouTube videos are the basis of our analysis. Specifically, we are going to extract the comment section of each selected video. The reasoning for designating YouTube as the ideal data source is threefold. Firstly, while there are many other video hosting sites (Vimeo, MetaCafe, Dailymotion), YouTube is the most popular, incorporating an innovative, interactive, and intuitive comment feature. Secondly, there exists a vast comment section for each video. Lastly, YouTube's underlying API allows interaction without requiring special permissions or complex setups.

Data Collection

We utilized random sampling to collect the data. One reason random sampling is beneficial is because of projects involving large datasets. When working with a project that exposes us to overwhelming amounts of data, it is more manageable to take a subset of that data. Another benefit of this method is its unbiased nature. By statistical principles, random sampling selects the desired data with equal chance. In other words, the method by which the data is collected does not influence the results of the project.

In order to manage the overwhelming amount of videos that exist in YouTube, we used random sampling to collect the videos in an unbiased manner. This way, the videos that we select will not influence the potential comments to be analyzed. We targeted YouTube's most popular videos for three reasons. First, in order to solve our client's issue of determining aspects of positive viral videos, we need to work with popular videos, as viral videos are popular by nature. Moreover, these videos, by definition, have surpassed the proprietary metrics YouTube uses to

designate videos as popular. Finally, it is more likely that the most popular videos have sufficient comments.

Of this popular video population, we decided to randomly select two videos for our random sample. By selecting two videos, we will create a baseline for our analysis which will include the range of emotions that each popular video invokes in its comment section. This is an important assumption of the videos that we expect, which is that overall each video will have a range of emotions and not have a purely negative or positive comment section. The range of emotions will skew to a positive or negative emotional range and understanding this pattern is one of the core tasks of our analysis.

In order to have access to YouTube's most popular videos, we consulted the website's documentation, which described aspects of YouTube's Data API¹. YouTube's Data API is a versatile input mechanism that has a vast library that allows us to interact with the underlying programming of YouTube. Specifically, we used YouTube's Data API (v3) to retrieve our population of popular videos, and then we used YouTube's Data API (v2) to retrieve the user comments for each video².

After learning how to use the particular section of YouTube's Data API (v3) to access YouTube's most popular videos, we used R code to query that section. Our R script was comprised of two parts. The first part focused on retrieving a list of popular videos and randomly selecting two. The second part enabled extraction of each video's comments. We generated a list of the fifty most popular videos from the US region. From this list, we then randomly selected

¹ "YouTube Data API (v3)." *Google Developers*. Google, n.d. Web. 08 Apr. 2015. <<https://developers.google.com/YouTube/v3/>>.

² "Google Documents List API V2 Developer's Guide: Protocol." *Google Developers*. Google, n.d. Web. 08 Apr. 2015. <https://developers.google.com/google-apps/documents-list/v2/developers_guide_protocol>.

the two sample videos for our analysis. The code used to generate the list and select the videos can be found in the Appendix 1.0.

Once we had our two sample videos, we wrote R script to retrieve the comments for each video via YouTube's Data API (v2), as YouTube's Data API (v3) for comments is under development and does not provide this functionality yet. To collect enough comments we programmed YouTube's Data API (v2) to page through each video's comment URL on the JSON response and used this URL to retrieve the next set of comments. Before YouTube's Data API (v2) paged through to the next comment section we had to bind the new comments with the previous one.

After collecting sufficient comments, we wrote all the comments to a CSV file in the working directory. This process was run twice, once for each video, so we could create two sample datasets. The total number of comments collected for each video was 3274 and 5037 for the first and second videos, respectively. Different comment totals were collected for the following reason. The first video only had a total of 3274 comments at the time of our collection, and we capped the second video in order to maximize the limited resources we had for this project. The code used to the download the comments can be found in Appendix 1.1

Table 1 below describes the thirteen default metadata variables that were extracted from our two randomly selected YouTube videos and downloaded into our datasets. These variables include information about the video and the author of the comment, but the content variable is the most important for our analysis because it is the actual comment.

Table 1 *Variables of comment section of YouTube's videos*

Variable	Description
id	URL that retrieves a specific comment using Youtube API.
published	Date of publication
updated	Date of the last modification
category	URL used by API to categorize the content retrieved (comment).
title	Title of the comment
content	The content of the comment encoded UTF-16
link	The same function of id variable
author	The name of the author of the comment
AuthorUri	URL that retrieves author's metadata.
YtChannelId	The YouTube Id of the author's channel
YtGooglePlusUserId	The author's Google Plus user Id
YtReplyCount	Number of replies for a single comment
YtVideo	Id of the YouTube video where the comment was posted

Data Munging

Data munging involves preparing the dataset in order to make it easier to work with. In other words, we produced a more clean dataset that will run more smoothly with our tools. Data preparation is necessary to make this project feasible given our resources, while at the same time maintaining data integrity. We dropped all variables of the datasets except for the content variable. This enabled us to focus solely on the comments, as the other variables are irrelevant to our project. We also ran both the datasets through a `qdap` R package, introduced later in the analysis section, to clean and prepare them further. The script written for this can be found in Appendix 1.3. The following cleaning procedures are completed using the `qdap` package.

- Remove bracketed text so that unicode within comments can be removed.
- Perform general text cleaning to remove extra white spaces and other textual anomalies.
- Strip text of unwanted or escaped characters and remove leading or trailing whitespaces.
- Add space after a comma and replace incomplete sentence end marks.
- Comments with two or more sentences are considered as single sentences with multiple punctuations because splitting of comments into separate sentences makes them to be analysed as separate comments. This results in the word counts returned by the polarity function to denote the number of words in each comments instead of each sentences.

Analysis

Our project involved determining the range of emotions in a viral video. We analyzed the comments of each video to determine the sentiment of each comment. Our analysis involved both manual and automatic work. We began by manually gathering facts about the two popular videos selected our data collection. The principal features of the videos are depicted in Table 2.

Table 2 *Features of selected popular videos*

Video#	Video Name	YouTube Video ID	Published	Video Owner
1	10 Extremely BIZARRE Phobias People Actually Have!	ZY7fz_9kStQ	2/21/15	Matthew Santoro
2	Katy Perry - Roar (Official)	CevxZvSJLk8	9/5/13	KatyPerryVEVO

To get a sense of the comments we worked with, we dived into Exploratory Data Analysis (EDA). EDA allowed us to generate initial thoughts of how to best analyze the comments. We manually reviewed the first several comments of each video. We found that

whether a user's comment was positive or negative depended on the nature of the video. If a video was intended to invoke negative reactions, a user's negative reaction would be considered a positive outcome. Conversely, a video that expected positive reactions, which attained unwanted negative reactions, would result in a negative outcome.

Our first video was about phobias, including fear of holes. For this video, we found that it was expected for users to have negative reactions, as fear of holes invokes revulsion. For this particular video, negative reactions were considered a positive outcome. Therefore, a user whose comments represented their revulsion to the fear of holes are positive comments, while users' whose comments revealed no feelings are negative comments.

Our second video was a music video. Since the video's lyrics describe strong feelings of empowerment and self-esteem, we found that the video's intended reactions are positive. So, any negative comments would be a negative outcome, and any positive comments would be a positive outcome.

Our manual analysis has implications for our business problem. Our client wants us to find the sentiment of viral videos, as this will help improve their products' videos and improve customer feedback. In our manual analysis, we found that the type of outcome, positive or negative, depends on the expected reactions of the video. Our client must ensure it knows what type of reactions it prefers to invoke. For product videos that aim to highlight the strong points of the product, the company most likely shoots for positive reactions. In this case, positive reactions reflected in comments are a positive outcome. By understanding the nature of the product video and the intended reactions, we can identify which reactions would result in the desired positive viral outcome. These possibilities are illustrated in table 3.

Table 3 *Relation between intended reaction, actual reaction, and outcome*

Intended Reaction	Actual Reaction	Outcome
Positive	Positive	Positive
Positive	Negative	Negative
Negative	Positive	Negative
Negative	Negative	Positive

The first column of Table 3 shows the intended reaction of the video, the second column shows the actual reaction a viewer might have to the video, and the third column shows whether the intended reaction meets the actual reaction.

We automatically analyzed the data by applying two text analysis methods. The two methods were Text Frequency Analysis and Text Sentiment Analysis. Since our project involved analyzing viewers' reactions, we needed to determine the sentiment of each comment. Text Sentiment Analysis involves assigning a score to each word, as well as the overall sentence or phrase being analysed.³ Text Frequency Analysis involves determining the statistical summary of the data in order to clearly demonstrate to our client the language being used to talk about their video. By understanding the most frequently used words, the client will be able to correlate the strongly used words with the viewers' reactions. We selected tools to automatically perform our text analyses and produce visuals that allowed us to effectively communicate our findings to our client.

³ Pang, Bo, and Lillian Lee. "Opinion Mining and Sentiment Analysis." *Foundations and Trends® in Information Retrieval* 2.1–2 (2008): 1-135. Web. 8 Apr. 2015.

For the text frequency analysis we used a tool called Voyant Cirrus.⁴ This tool creates a word cloud that relates to the frequency of words that appear on a document. We additionally applied the stop word filter from Cirrus, which “contain so-called function words that don’t carry as much meaning, such as determiners and prepositions (in, to, from, etc.).⁵”

Text sentiment analysis is the method of assigning words with a score to identify them as positive, negative, or neutral. The technical word for this is polarity. Text sentiment analysis is a crucial part of our project as it scientifically allows us to analyze the value of the words in a mathematical format. For the text sentiment analysis we used an R package called qDap, which is defined in Appendix 3.0. Using this qdap package allowed us to run a polarity function that analyzes and weighs each sentence of the comments with a polarity (positive, negative, or neutral) score. This sort of advanced analysis provides a way to quantifiably connect the raw data and the emotional value we are looking for.

The above scoring functions in qdap make the following assumptions about data⁶ being processed:

- Each row contains a single sentence
- Each sentence contains only one end-mark
- Commas are followed by a space
- Numbers and symbols are unimportant (ignored)
- Words are spelled correctly and contain no escape characters

⁴ Sinclair, Stéfan, and Geoffrey Rockwell. Cirrus. Computer software. Voyant Tools: Reveal Your Texts. Vers. 1.0. Voyant Tools, n.d. Web. 08 Apr. 2015. <<http://voyant-tools.org/tool/Cirrus/>>.

⁵ Sinclair, Stéfan, and Geoffrey Rockwell. "STOPWORD LISTS." Voyant Tools Documentation. N.p., 2015. Web. 8 Apr. 2015. <<http://docs.voyant-tools.org/ui/stopwords/>>.

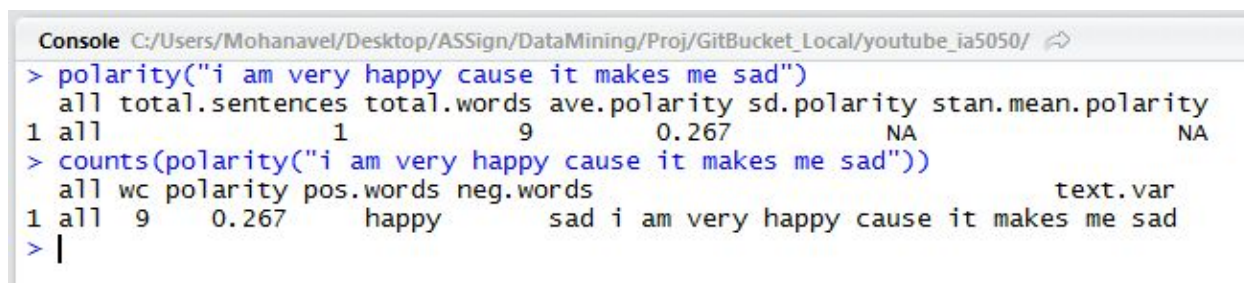
⁶ Rinker, Tyler W. "Cleaning Text and Debugging." (n.d.): n. pag. *Qdap Package Vignette*. 4 Oct. 2014. Web. 11 Apr. 2015. <http://cran.r-project.org/web/packages/qdap/vignettes/cleaning_and_debugging.pdf>

In order to take into consideration the aforementioned assumptions, our data munging process, detailed earlier, cleaned the data accordingly.

The Polarity Score (Sentiment Analysis) function provided by qdap package uses an algorithm which determines the polarity of a comment. The polarity score generated by this function is dependent upon the polarity dictionary used. The function's default dictionary by Hu & Liu (2004) is used for this analysis⁷.

The polarity function takes each comments as input and assigns them polarity score by tagging polarized words using the default dictionary. For example, as shown in the second command in Figure 1 below, words from the comments are tagged as either positive or negative based on the polarity dictionary. "happy" was labeled under pos.words, while "sad" was labeled under neg.words.

Figure 1: *Tagging of polarized words and sample output of polarity function*



```

Console C:/Users/Mohanavel/Desktop/ASSign/DataMining/Proj/GitBucket_Local/youtube_ia5050/
> polarity("i am very happy cause it makes me sad")
  all total.sentences total.words ave.polarity sd.polarity stan.mean.polarity
1 all           1           9      0.267          NA              NA
> counts(polarity("i am very happy cause it makes me sad"))
  all wc polarity pos.words neg.words          text.var
1 all  9   0.267   happy      sad i am very happy cause it makes me sad
> |

```

The words that are not found in both positive and negative dictionary files are considered as neutral. These neutral words holds no value other than affecting the word count. Then, (default) four words before and two words after tagged polarized words are clustered together as valence shifters. The words in this cluster are tagged as neutral, negator, amplifier, or de-amplifier. However, only the words found after the comma will be considered if a cluster of

⁷"Opinion Mining, Sentiment Analysis, Opinion Extraction." N.p., n.d. Web. 11 Apr. 2015.
<<http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html>>

words contains a comma before the tagged polarized word. “Each polarized word is then weighted and further scored by the number and position of the valence shifters directly surrounding the positive or negative word.⁸” Finally, these scores are summed and divided by the square root of the word count, thereby resulting in an unbounded average polarity score as shown in the first command in the Figure 1. The script written to perform these qDap functions can be found in Appendix 1.4.

Results

The purpose of this section is to discuss our findings of the automatic analysis. The word cloud is a visual representation of the text frequency analysis. In this case, the larger the word in the cloud, the more frequent it is used throughout all the comments in the dataset. Additional text frequency analysis can be found in Appendix 2.0.

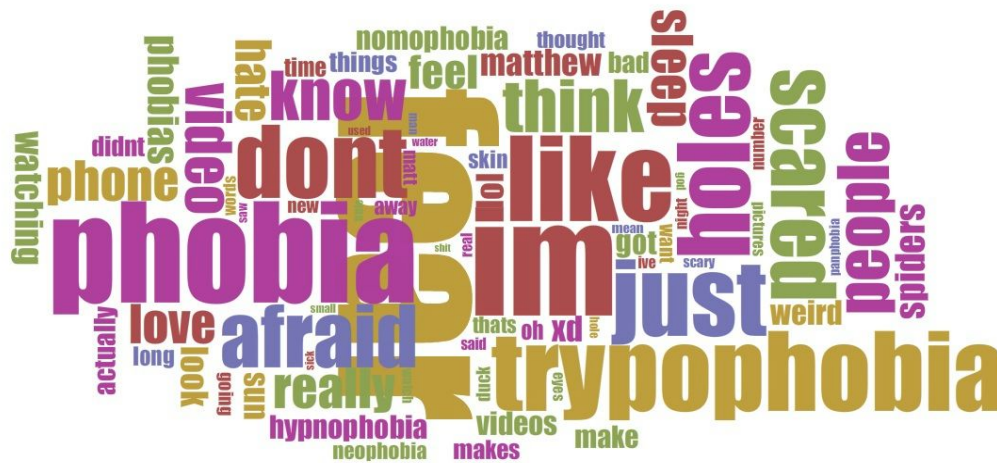


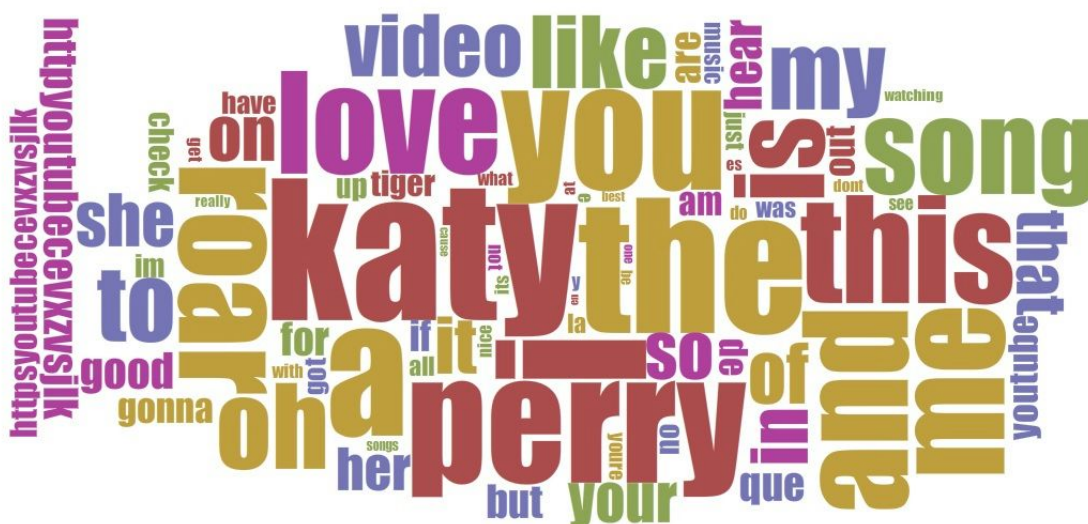
Figure 2: *Word cloud for comments of phobia video*

⁸ Rinker, Tyler W. "Qdap Package Vignette." N.p., n.d. Web. 11 Apr. 2015. <http://trinker.github.io/qdap/vignettes/qdap_vignette.html#polarity>

From Figure 2, we found that “fear,” “phobia,” and “like” are among the biggest words in the word cloud. This means that words like “fear” and “phobia” were very frequent. Clearly, the Phobia video sparked viewers to have conversations about the subject matter of the video.

The Music video was a Katy Perry music video. From figure 3 we found that “Katy”, “Perry”, “Love”, and “Song” are among the biggest words in the word cloud. The conversation sparked by this video shows that the viewers were really focused on the artist and the song.

Figure 3: *Word cloud for comments of Music video*



For our client, we can communicate that the results of the word cloud will depict the words commented about most. Based on our findings, it is highly likely that these words will relate to the most highlighted features of the product.

Taking the analysis deeper than the word cloud, the next step was sentiment analysis of the comments with the qDap functions. Here, we show the top 5 negative and positive

comments. For the Phobia video the top 5 negative comments ranged from -1.343 to -1.878. The top 5 positive comments ranged from 1.000 to 1.533.

Figure 4: *Top 5 comments with most negative and positive polarity for phobia video*

```

Console R Markdown x
C:/Users/Mohanavel/Desktop/ASSign/DataMining/Proj/GitBucket_Local/youtube_ia5050/
> #display Top 5 comments with most negative polarity for Video 1
> datcom1[1:5,c(3,6)]
  polarity
text.var
118 -1.878
i have severe general anxiety disorder
2302 -1.652
ly strangest phobia of all must be if you are phobophobic a fear of fear come on and it is actually a real phobia
1548 -1.605 i dont know if this is a phobia but the high pitched alarm noise hurts my ears and i get really scared i don
t mean the ringing i mean those newer ones that make a high pitched squeak they kind of make me deaf for a second
597 -1.441
since those who have this fear fear everything do they fear the fear of everything the answer i
s yes a fear is something and since people with this fear fear everything they are afraid of even their own fears
162 -1.343
i looked up some of these fears and the fear of losing your phone is treated more seriously
than fear of holes but the fear of holes is older than the fear of losing your phone even though that is younger
> #display Top 5 comments with most positive polarity for Video 1
> datcom1[nrow(datcom1):(nrow(datcom1)-5),c(3,6)]
  polarity text.var
130 1.533 awesome podcast matthew really good my son loves it
418 1.500 good work like it
570 1.470 you really love the plush emoji
2418 1.155 wow just wow
2154 1.155 wow i wow
2729 1.000 i love your work
> |

```

From figure 4, when we compare the positive and negative comments to each other, we can see how the negative comments have higher ratings. This shows that the top reactions overall were negative, which falls in line with the theme of the Phobia video.

For the Music video the top 5 negative comments ranged from -1.207 to -2.121. The top 5 positive comments ranged from 1.732 to 3.867.

Figure 5: *Top 5 comments with most negative and positive polarity for music video*

```

Console R Markdown x
C:/Users/Mohanavel/Desktop/ASSign/DataMining/Proj/GitBucket_Local/youtube_ia5050/
> #display Top 5 comments with most negative polarity for Video 2
> datcom2[1:5,c(3,6)]
  polarity text.var
996 -2.121 fake fake fake fake fake fake views reported
1434 -1.443 fuck you katy perry you stupid and bitch fuck you fuck you
3474 -1.414 fuck haters
1968 -1.342 this bitch is trash dislike
2884 -1.207 o ya and people dont be afraid to be you stand high be weird crazy silly dont change for others
> #display Top 5 comments with most positive polarity for Video 2
> datcom2[(nrow(datcom2)-1):(nrow(datcom2)-6),c(3,6)]
  polarity text.var
1983 3.867 love love love great song great vid great voice
2864 2.123 very love you very very good
1686 2.078 good very good
366 1.789 good good good im like
3248 1.732 well well well
1187 1.732 love love love

```


From figure 5 we can see how the top positive comment resulted in with a score of 3.867, this comment used the word “love” and “great” 3 times each. By comparing the average sentiment score of both positive and negative comments we see that the average positive score is 2.664 whereas the average negative score is 1.505. Next, we will look at the overall polarity scores and see if this trend holds true.

Figure 6 below shows the polarity distribution for the phobia video, where each dot represents a comment. The black line depicts the average score, and the thick red line designates neutral comments having no value and score of 0. The bulk of the comments are negative.

Figure 6: *Polarity distribution of comments for phobia video*

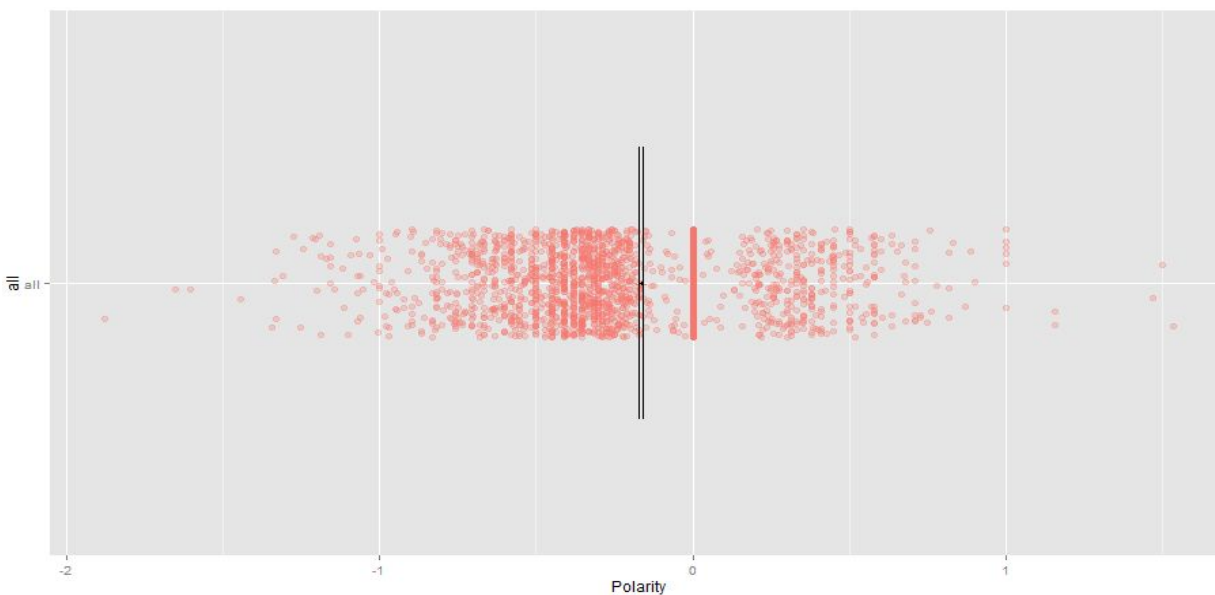


Figure 7: *Polarity statistics of all comments for phobia video*

```

Console C:/Users/Mohanavel/Desktop/ASSign/DataMining/Proj/GitBucket_Local/youtube_ia5050/
> poldat1
all total.sentences total.words ave.polarity sd.polarity stan.mean.polarity
1 all 2817 42795 -0.164 0.354 -0.462
> |

```


Figure 7 is the actual overall data that the polarity heat map is based on. Included in this figure is the Standard Mean Polarity, which excludes all the neutral comments. This score of -0.462 further shows that the overall sentiment for the Phobia video is negative.

In figure 8, the polarity Music video heat map shows how overall the comments are positive. On Top of which they are densely clustered together that they almost form a solid red block.

Figure 8: *Polarity distribution of comments for music video*

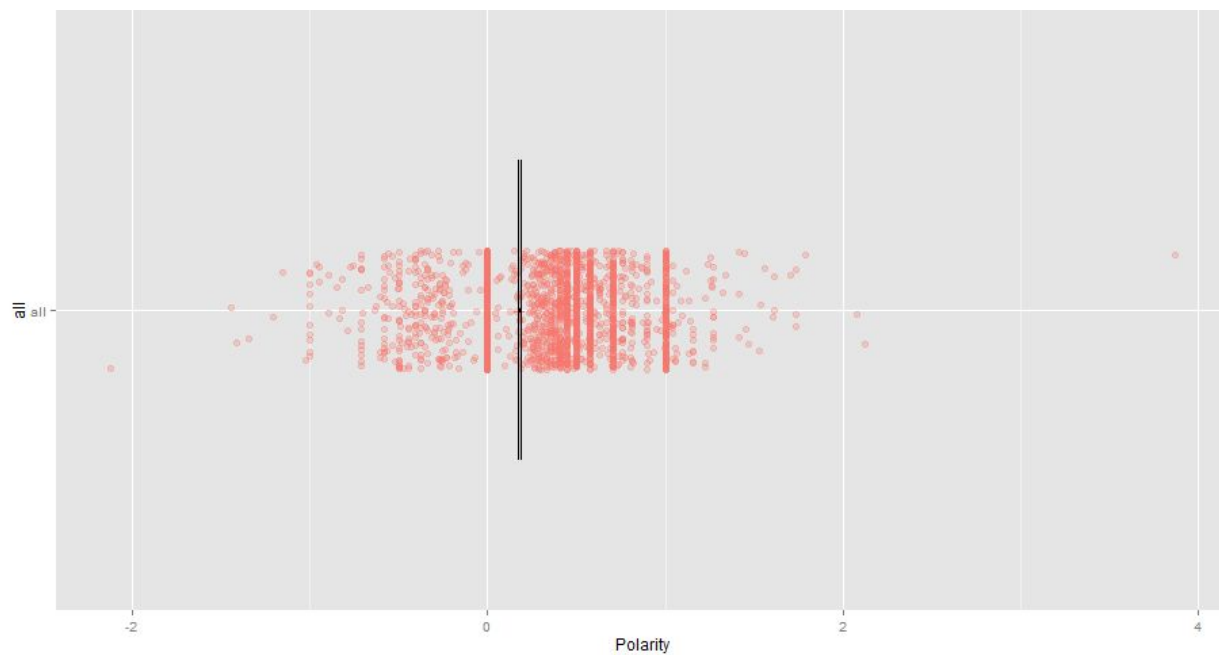


Figure 9: *Polarity statistics of all comments for music video*

```

Console C:/Users/Mohanavel/Desktop/ASSign/DataMining/Proj/GitBucket_Local/youtube_ia5050/
> pol1dat2
all total.sentences total.words ave.polarity sd.polarity stan.mean.polarity
1 all 4346 36766 0.179 0.37 0.485
> |

```

Figure 9 is the overall data points of the Music video heat map. Just as the Phobia video, when we look at the Standard Mean Polarity we can confirm that with a score of 0.485 this video has an overall positive reaction with viewers.

After interpreting the results from each tool, we drew final conclusions about each video. We found that the Phobia video was mostly negative on the spectrum. Because the video intended on garnering negative reactions, this implies that the reactions were a success. Given our word cloud, we find that the frequent words, such as fear, phobia, and scared, have to do with the intended reactions. Therefore, it could be possible that these particular words more strongly influence the heavy negative reactions. Similarly, for the music video the frequent words such as love, roar etc. have to do with the intended reactions. Therefore, it could be possible that these words more strongly influence the substantial positive reactions. Thus the intended reactions for the music video aligns strongly with the actual reaction and is considered to be a success.

Risk Assessment

The purpose of this section is to highlight risks we faced, as well as how we accepted or mitigated those risks. From as early as our data collection phase, we were hit with two risks; one involving video selection and the other involving number of comments. While we knew that our analysis needed to be focused on popular videos, we needed to ensure proper selection of those videos. The risk we faced was trying to select 2 popular videos from amongst the millions of YouTube videos in an unbiased manner, since our conscious and unconscious preferences could influence our selection.

Part of the criteria for our video selection was selecting 2 videos that had sufficient comments for us to perform a meaningful analysis. The risk here is twofold. First, too little comments will not be enough data to get a true sense of the emotion that a video provokes. Second, working with YouTube comments has potential to expose us to millions of comments. Too many comments would overwhelm the limited resources we had to solve our business problem.

We wrote a script to mitigate these first two risks. For video selection, we used R to write a random sample algorithm to query Youtube using its API which inturn returns URLs of random videos with a specific set of conditions. To counter the aforementioned comments risk, we decided to create datasets that consisted of a minimum of three thousand comments and maximum of five thousand comments each.

While wrangling the data in preparation for our analysis, the overall risks involved cleaning the data appropriately. We had to strike a balance with cleaning the data without altering its integrity. Specifically, we encountered the following: misspelled words, spam that plagued the comment section, and unicode. We determined that cleaning the data too much, i.e. correcting misspelled words or sentence structure, risked altering the text sentiment analysis. If we did not eliminate the unicode, then our word frequency analysis would be skewed as well. However, mitigating the unicode risk forced us to accept the risk that actual information within brackets are removed. Based on our observation, most users rarely utilize brackets for expressing useful information; instead, they use it to mimic emoticons. This presents minimal risk on our analysis, thus justifying it's acceptance.

Table 4 *Project risk and controls*

Risk	Controls	Remaining Risk
YouTube's video population is huge.	Use YouTube API to retrieve most popular videos and random sampling.	The method used by YouTube to designate popular videos can be biased by YouTube's customers (advertisement) or unknown interests.
Too little comments	Allow more time to obtain enough comments. Daily surveillance of the comment section of the video.	Project deadline is a constraint.
Too much comments	Limit the number of comments to 3000.	A big amount of spam comments in the sample.
Comment's text encoded (UTF-16)	Eliminate UTF-16 tags and special characters	Lose of information, i.e. emoticons, non standard characters from other languages.

Conclusion

We concluded our data mining project by developing a comprehensive methodology. The word cloud and polarity distribution allows us to analyze any client's video and produce results that can be interpreted similar to the two videos in our project. Our final methodology is as follows:

1. Attain Client's Product Video
2. Elicit Video's Intended Reactions
3. Generate Word Cloud
4. Generate Polarity Distribution
5. Determine Actual Reactions
6. Interpret Final Outcome

Since our analysis of the two videos in the project were representative of the popular video population, we can be sure that any popular video can go through our methodology and be interpreted accordingly. After applying our methodology on a client's video, we draw conclusions that provide insight for how the audience reacts to the client's product. These conclusions can tell the client whether the video was positively or negatively popular as well as how its video invoked the positive or negative reactions.