IS434: Social Analytics and Applications

Term Application Project: Final Report

Section G1
Group 13

Vanessa Goh Ze Hui
Wu Yunheng (Winston)

Date of Submission: 23 November 2017

About Ministry of Funny

Ministry of Funny is a local company which produces original videos that combine humour and technology to tell stories, challenge stereotypes and tackle everyday issues.

They have substantial social media presence on almost all of the common social media platforms, such as Facebook, YouTube, Twitter and Instagram. They have approximately 44,000 likes on Facebook, over 100,000 subscribers on their YouTube channel, 500 followers on Twitter and over 6,000 followers on Instagram.

Currently, they only review their analytics occasionally, but are unable to gain any significant insights. More often than not, they decide to do videos based on their "gut" feelings by leveraging on data-driven insights to guide their decisions.

Business problem

Overall Problem: Ministry of Funny's main problem revolves around having a lack of in-depth knowledge about their audiences across their various social media platforms, what type of content appeals to them, and what type of content repels them, and it has been more like a hit-or-miss in the type of videos they are doing now. They would like to focus on whether their videos have a tangible impact on their viewers in terms of having a "social impact", in which their viewers are more inclined to engage in social conversations about taboo topics.

Therefore, Ministry of Funny has expressed interest in tracking whether different video categories receive more attention on different social media platforms, specifically on Facebook and YouTube. Examples of various video categories that Ministry of Funny has are prank videos, skits about social phenomenons in Singapore and parody videos. Beyond focusing on positive/negative sentiments for videos, Ministry of Funny also stated that they want to focus more on videos by themes that are more edgy, and look for sentiments in terms of how much a social conversation it starts, like in terms of how thought-provoking it is. For example, they currently have 3 videos under the casual racism theme, and will have more taboo themes for videos in future to generate more social conversation.

Data Collection

We will be using a combination of data that we have obtained from our client and data that we have extracted using the various APIs. Descriptions of the data we received from our client and data that we are extracting is as follows:

Data Received:

● YouTube API Key, client_id.json file (downloaded from console), exported reports for the months of July, August, September and October (month by month), OAuth 2.0 token, YouTube V3 Discovery Document
● Assigned YouTube Roles - Editor
● Assigned Facebook Page Roles - Analysts

Data Extracted:

- Extracted all Ministry of Funny's video comments and titles from Ministry of Funny's YouTube channel via YouTube Data API
- Extracted all Ministry of Funny Facebook Page likes, comments, reactions, replies (Facebook Graph API)

Using the Facebook Graph API, all posts data on the Ministry of Funny Facebook Page were first extracted. This includes information such as its unique ID, the time it was published, and more importantly, the likes and reactions that pertained to each post. These were saved in a first CSV file. Subsequently, all comments that pertained to each post were then retrieved, including the published time and the contents of the comment, and was saved in a second CSV file. Following that, we insert these records into our AWS mySQL database instance for longer-term storage.

- YouTube video details from the playlist uploads (YouTube Data V3 API)
- Comments on individual YouTube videos (YouTube Data V3 API)

With reference to the YouTube Data API v3, we obtained the full sample code in Python that was on the API, for the data we wanted to obtain. We then debugged the sample code as it was not the most updated, and in the process, configured the settings needed for YouTube OAuth 2.0 authorisation. These were done on Google's developer console and Google developers platform. Clear instructions on the necessary configurations to be made were not given on the API, so we had to research and figure out the changes to be made on our own. For instance, the dependency package for apiclient should now be googleapiclient. We also had to set up and generate a refresh token that would continually renew when they expire in order to be granted access to the data in the long-term. In order to get all configurations and information at one go from our client, we tested out the above on a personal YouTube account first and scheduled a meeting soon with our client to extract their actual data.

However, the extracted data from our client's YouTube console did not include individual video comments, and therefore, we had to retrieve the video comments via

the YouTube Data API eventually. Subsequently, we also inserted these comments into the MySQL database on AWS for easier storage and retrieval as well.

Data Cleaning/Transformation

Using python scripts to invoke the Facebook Graph API, the data that was returned was originally in JSON format. However, we converted all the data into specific columns that pertained to each field (e.g. num_reactions, num_likes, etc.), and saved them in a CSV file. This was done in two separate stages - a CSV file with the posts data and a second CSV file that had the comments data.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	status_id	status_message	link_name	status_typ	status_lin	permalink_url	status_published	num_reac	num_com	num_shar	num_like	num_love	num_wov	num_haha	num_sads	num_angrys
2	16682905(	NEW VID! Ever	Ownself C	video	https://w	https://www.fa	6/10/2017 2:16	83	3	29	70	3	0	10	0	0
3	16682905(	The FIRST	Ownself C	video	https://w	https://www.fa	5/10/2017 6:45	13	0	3	12	1	0	0	0	0
4	16682905(	Reading this	HBO Asia I	link	http://ww	https://www.fa	16/9/2017 23:52	179	5	23	167	12	0	0	0	0
5	16682905(	WE'VE GOT AN	We've got	video	https://w	https://www.fa	15/9/2017 5:58	76	2	9	69	5	2	0	0	0
6	16682905(	T minus 6 days	Honesty n	link	http://m.t	https://www.fa	10/9/2017 23:26	29	2	2	24	5	0	0	0	0
7	16682905(	Singaporeans	Ministry C	photo	https://w	https://www.fa	3/9/2017 1:20	22	0	2	14	0	0	8	0	0

Snapshot of Extracted Facebook Dataset (Posts)

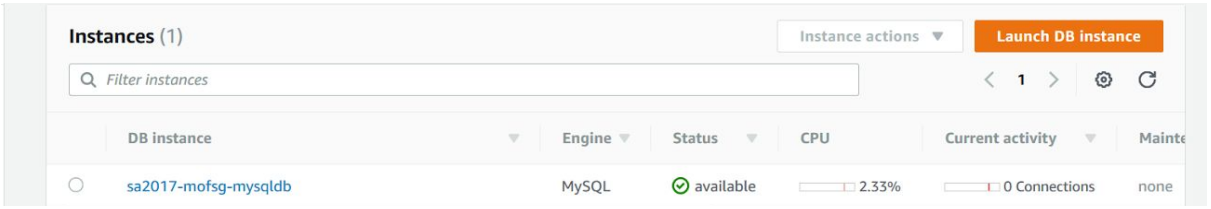
	A	B	C	D	E	F	G	H
1	comment_	status_id	parent_id	comment_message	comment_author	comment_published	comment_	status_published
2	16009559	166829056716989_		Hmmm..I seriously think that their videos	Cyprian Justin Lee	6/10/2017 11:55	3	6/10/2017 2:16
3	16009559	166829056716989_		Good video and production all ,but .....do	Jeff Tan	7/10/2017 19:05	1	6/10/2017 2:16
4	16009559	1668290	16009559	I Agree anyone else?	Kelapa Sawit	13/10/2017 6:00	0	6/10/2017 2:16
5	16009559	166829056716989_		best fan !	Kelapa Sawit	13/10/2017 6:00	0	6/10/2017 2:16
6	15839941	166829056716989_		Can! Omg I lol watching the trailer	Hossan Leong	17/9/2017 2:51	2	16/9/2017 23:52
7	15839941	166829056716989_		It's gonna be so good Haresh!	HeeJung Foo	17/9/2017 5:35	2	16/9/2017 23:52
8	15839941	1668290	15839941	Anyone else agree?	Kelapa Sawit	2/10/2017 4:00	0	16/9/2017 23:52

Snapshot of Extracted Facebook Dataset (Comments)

Furthermore, the structure of the data was arranged such that each comment would be matched with its corresponding Facebook post (i.e. the post that the user commented on) or with its parent comment ID (i.e. the top-level comment that another user replied to). This was crucial in extracting the time differences between the Facebook post and when the comment was made, which aided in the preparation of the buzz analysis in the subsequent section.

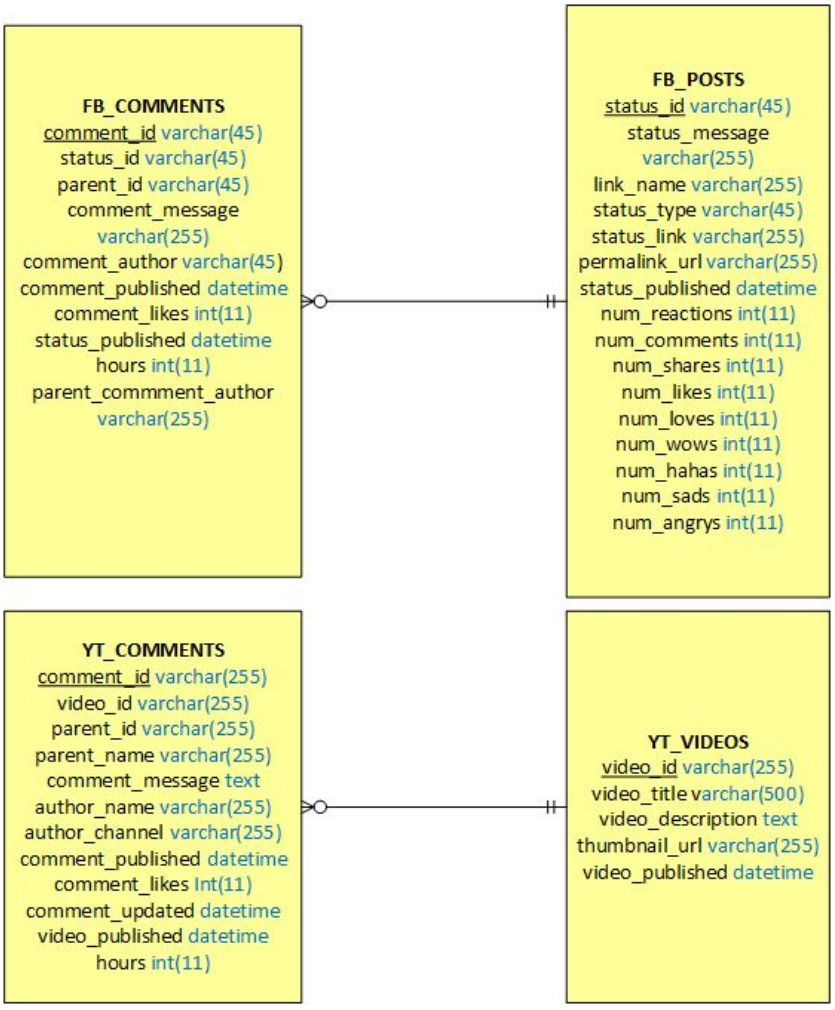
Additionally, beyond using stop words and porter stemming for our analyses, we also incorporated the use of a wordlist as a form of corpus to remove gibberish and non-English words from our Frequency Analysis, so as to provide a more accurate overview of the keywords or buzzwords being said.

Following the midterm presentation, we have followed through with our plan to store the crawled data on an instance of the MySQL relational database on AWS, which will facilitate easier data retrieval and analysis in future.



Screenshot of AWS RDS Instance

Additionally, we have come up with a first version of a logical ER diagram showing the various entities pertaining to the data that we obtained from Facebook and YouTube, along with their attributes and how they will be captured within the database.



Logical ER Diagram of Captured Data

Assumptions/Limitations

The following points are the assumptions or limitations that we have envisioned at this stage of the project.

1. After using the LDA Topic Modelling for Facebook comments, the output for the various topics were deemed to be of little use to our client. This may be because their videos target a widespread array of topics, resulting in insufficient data to generate meaningful clusters or topics as a result.

2. A future point of expansion would be to incorporate the use of a Singlish corpus to capture Singlish sentiment analysis. As there is no pre-existing Singlish corpus, Singlish words would have to be gradually added over time in order to build up the Singlish corpus, so that sentiments from Singlish comments can be factored in too. This would help both the Frequency Analysis and the Sentiment Analysis to have an even higher rate of accuracy.

3. Additionally, we were not able to analyse emojis such as 😁 this time, and were only able to analyse emoticons such as :) and :D. However in the future, it will further enhance the degree of accuracy for the analysis, if emojis can be factored in as well.

4. In addition, we have not made the dashboard completely dynamic yet, in terms of the corresponding analyses and visualisations showing when our client chooses a specific post or video. The backend SQL queries have been set up for this purpose, allowing us to query for individual post/video visualisations below. We will work on this front-end integration before handing over the finished dashboard to our client.

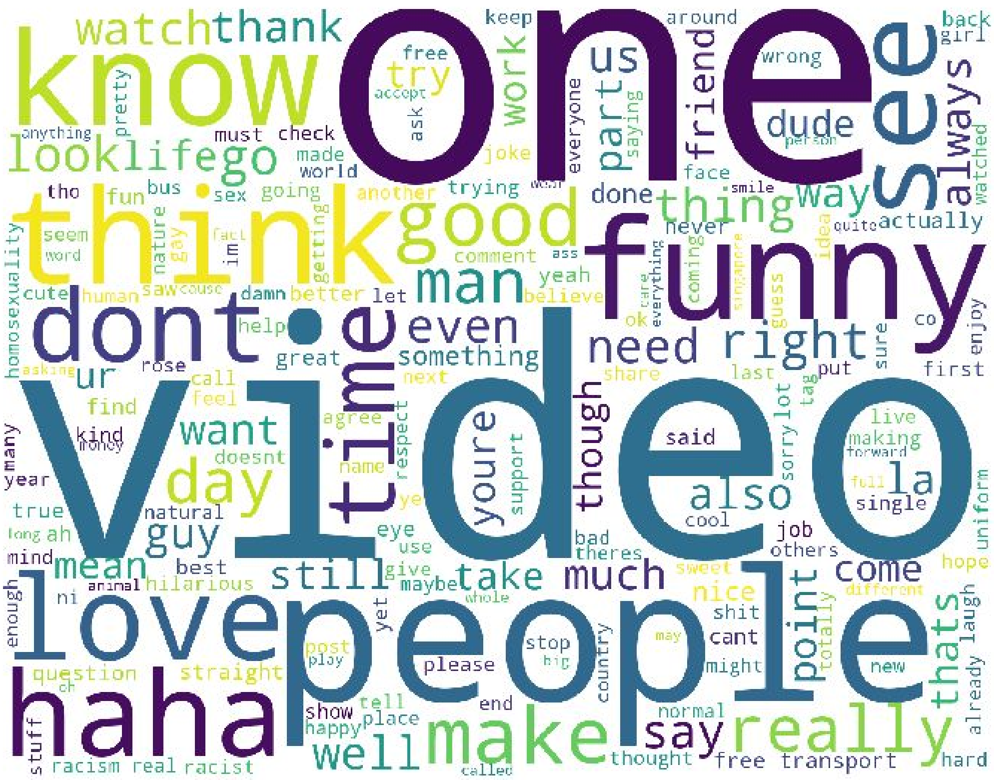
Technologies/Tools/Methodologies

The table below shows the summary of the technologies, tools and methodologies that we have used and are planning to use for our project:

Task	Technology/Tools/Methodologies
Facebook Data Extraction	Facebook Graph API
YouTube Data Extraction	YouTube Data V3 API
Database Storage & Retrieval & Dashboard	MySQL, Python Flask
Frequency Analysis - WordCloud	Python Script
Facebook Buzz Analysis	Python Script
Survival Analysis	Kaplan-Meier Estimate + Python Script
Sentiment Analysis	Python Script + VADER package
Clustering Analysis - LDA Topic Modelling	Python Script + gensim package
Social Network Analysis - Influencers	Python Script

Frequency Analysis - WordCloud (Overall - Facebook)

Following our midterm presentation, we further cleaned the comments we extracted from both Facebook and YouTube comments. This was done with the help of a wordlist, which helped to remove gibberish words from the WordCloud and only ensured that proper English words would be visualised. Based on all of the comments that were extracted from Facebook, we visualised a WordCloud that depicts the most frequently used words in our client's Facebook comments.



WordCloud Visualization for Facebook Comments

As seen above, the most frequently used words mostly reflect appreciation of Ministry of Funny's humour in the Facebook posts through words such as "funny" and "haha". With the wordlist, it helped to remove variations of "haha" that were present in the previous WordCloud, and other gibberish from the WordCloud as well.

Frequency Analysis - WordCloud (Individual - Facebook)

Additionally, we also made it possible for the client to generate WordClouds on a per-post basis (Facebook). As the nature of their videos belong to a myriad of themes, the ability to have an overview of what comments people are stating will allow our client to better assess if their videos have allowed them to reach their intended goals of sparking off conversations among viewers, especially for videos that have conventionally taboo themes.

For example, one particular video was a parody about the practicality behind the day in which there was free transport for all National Service personnel, but only if they were in uniform, to commemorate 50 years of service (Video link: https://www.youtube.com/watch?v=fJDF6PIkZlg). The generated WordCloud of all pertaining Facebook comments in relation to this video is as follows:

WordCloud Visualization for Facebook Post on Free Transport for NS50

For this post-specific WordCloud, this gives our client a more accurate view of what type of keywords or buzzwords are commonly repeated, and whether they have managed to spark off a pertinent conversation among their viewers.

Frequency Analysis - WordCloud (Overall - YouTube)

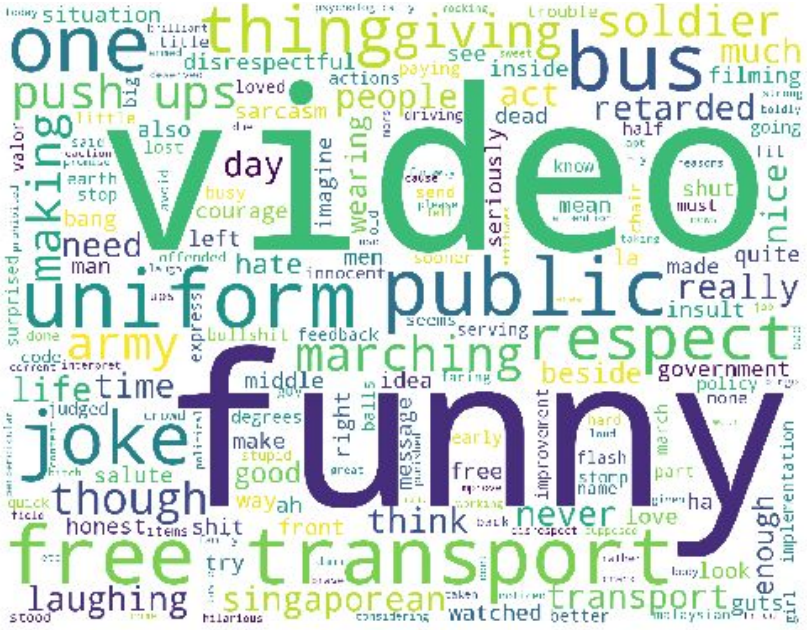
We also performed a frequency analysis for the overall YouTube video comments, as shown in the WordCloud below.

WordCloud Visualization for YouTube Video Comments

In comparison with the overall frequency analysis for Facebook, there are common recurring words in both WordClouds, such as ‘funny’. This may indicate that there may be common viewers across both platforms that are also commenting actively on Facebook and YouTube, resulting in similar words being used.

Frequency Analysis - WordCloud (Individual - YouTube)

In comparison with the individual Facebook Frequency Analysis, we also performed a frequency analysis on the individual video that the Facebook post was referring to.



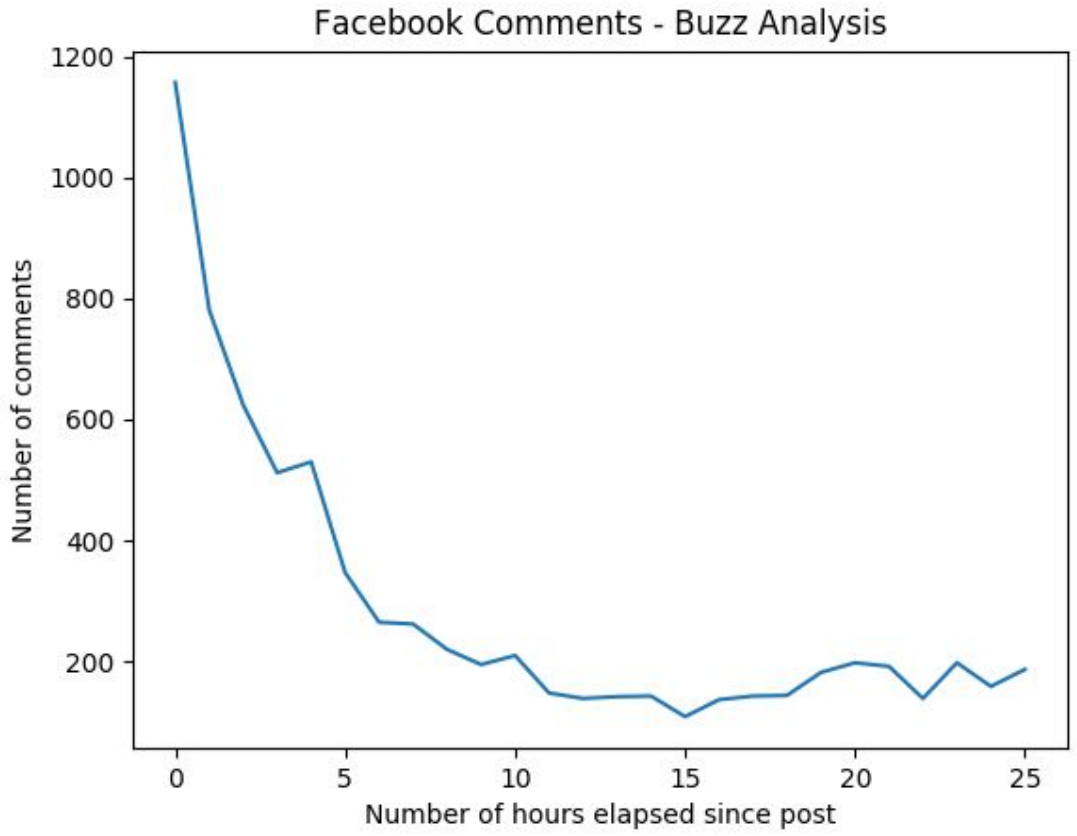
WordCloud Visualization for YouTube video on Free Transport for NS50

For this video-specific WordCloud, it is more evident that most viewers on YouTube found the video funny, and there was higher frequency of certain keywords in comparison with the Facebook video. Therefore, this gives our client a more accurate view of what type of keywords or buzzwords are commonly repeated, and whether they have managed to spark off a pertinent conversation among their viewers.

Buzz Analysis (Overall - Facebook)

We also conducted a buzz analysis for Ministry of Funny's Facebook data, which looked at the distribution of comments that happened within the first 24 hours of a Facebook post in order to ascertain how much conversation on the video's theme was generated.

.



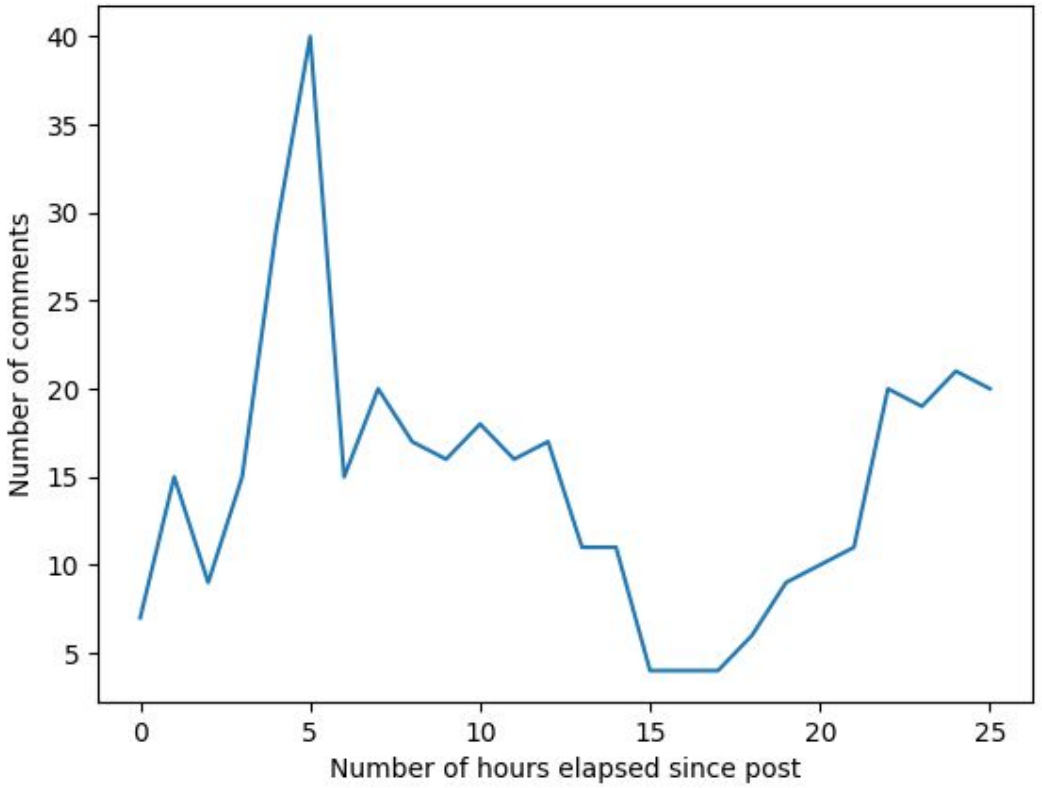
Line graph depicting number of comments against number of hours elapsed since a Facebook Post by the administrator

Nearly half of the slightly over 14,000 total comments that we extracted had occurred within the first 24 hours of a post, with an overwhelming number of comments (approximately 1,150) occurring within the first hour of the post, before tapering off sharply and gradually plateauing as the number of hours increased.

Buzz Analysis (Individual - Facebook)

Similar to the frequency analysis, we also included an additional function that will allow our client to choose a specific YouTube video or Facebook post, in order to see the buzz analysis for it.

For example, one of their Facebook posts that garnered a high number of comments was on theme of being LGBT and the questions that LGBT people often get from straight people (Video link: https://www.youtube.com/watch?v=GmDCoLKtDQk).



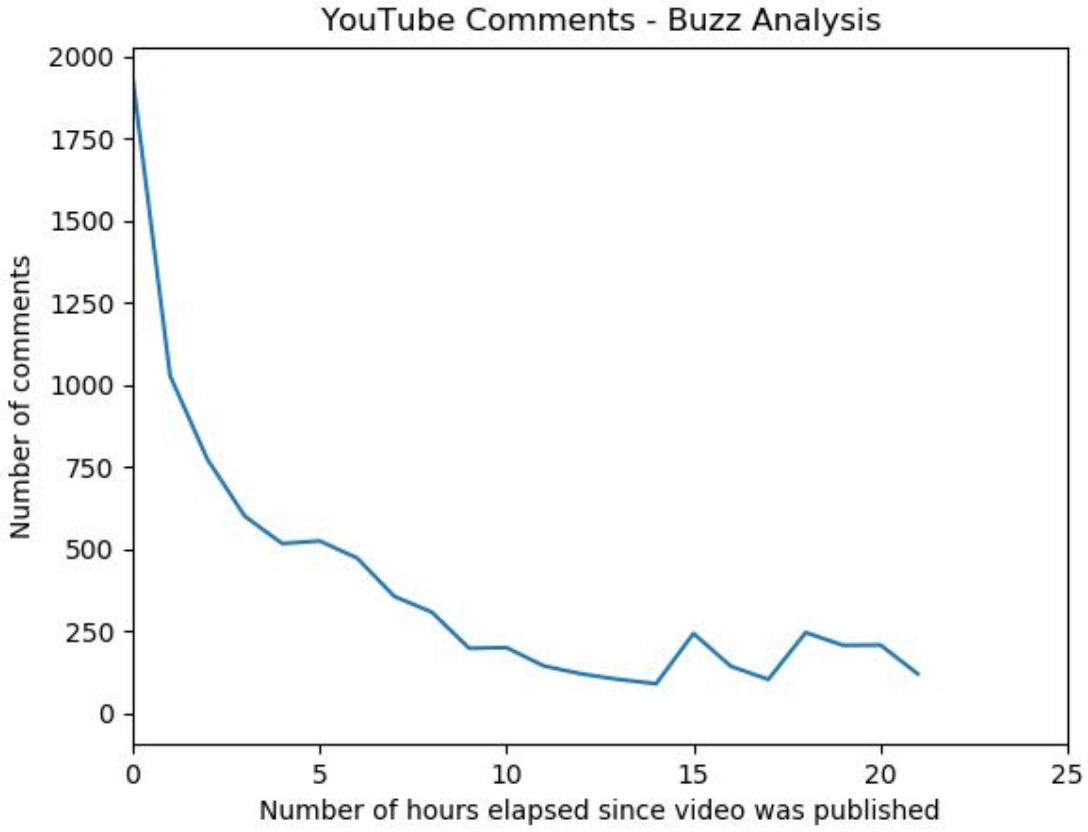
Buzz Analysis for a Facebook Post on questions that LGBT people get

The buzz analysis for this post is particularly noteworthy as it plateaus at the fifteenth hour mark, but then spikes up again gradually. This could be meaningful for our client as they are better able to pinpoint what may have caused a spike in a plateauing buzz analysis, such as a theme that is mentioned in mainstream media that garners more interest than usual on a particular topic.

Buzz Analysis (Overall - YouTube)

Also, we conducted a buzz analysis for Ministry of Funny's YouTube data, which looked at the distribution of comments that happened within the first 24 hours of a YouTube video being uploaded in order to ascertain how much conversation the video has generated.

.



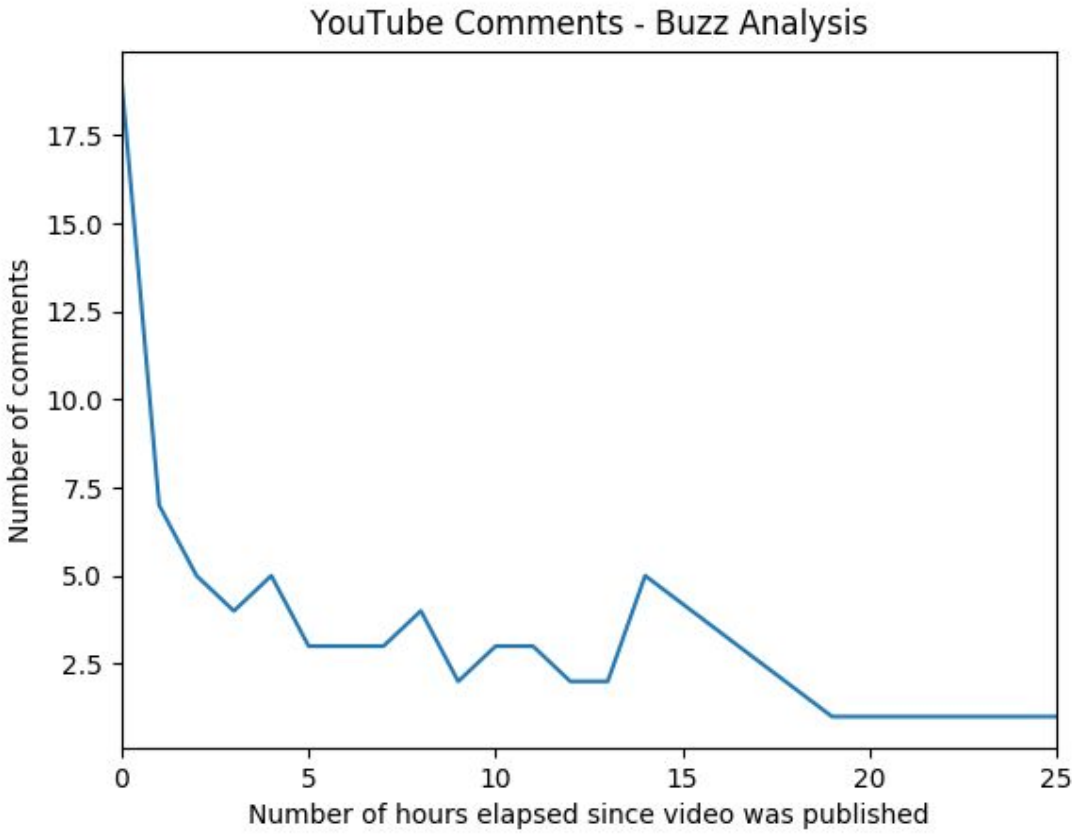
Line graph depicting number of comments against number of hours elapsed since a YouTube Video upload by the administrator

In line with the buzz analysis for Facebook, the vast majority of the total comments that we extracted from YouTube had occurred within the first 10 hours of a video upload, with an overwhelming number of comments (approximately 2,000) occurring within the first hour of the post, before tapering off gradually.

Buzz Analysis (Individual - YouTube)

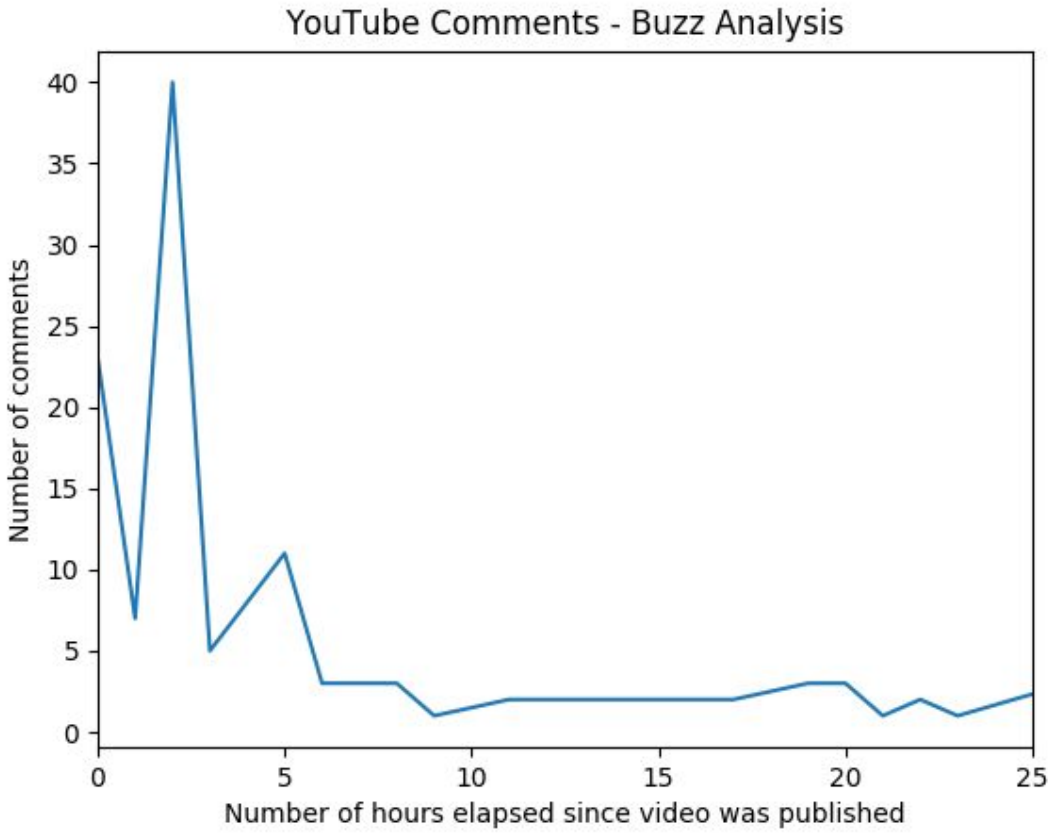
We also provided the option for our client to perform a buzz analysis on a per-video basis.

Again, analysing the video comments for the above LGBT video used for the individual Facebook Buzz Analysis, the graph below was derived.



Buzz Analysis for a YouTube video on questions that LGBT people get

This buzz analysis shows that for this particular YouTube video, the trend of the volume of video comments trailed downwards after posting, even though it received a slight hike in comments at around the thirteenth hour mark. However, it eventually plateaued at the twentieth hour mark.



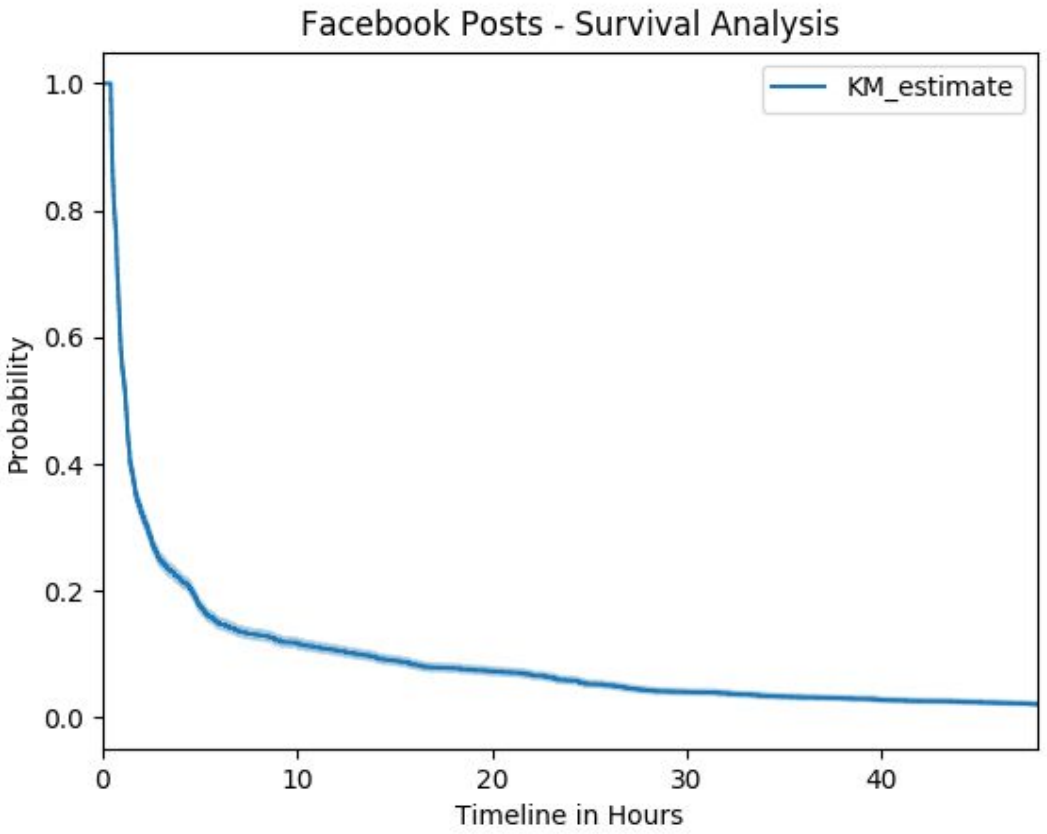
Buzz Analysis for a YouTube Video that was sponsored by Uber

Furthermore, a buzz analysis for sponsored or paid videos will help Ministry of Funny to have a more measurable means of tracking the attention garnered, which they can show to their own clients as well. A sponsored video with a long survival analysis would allow Ministry of Funny to command higher fees as a result of more leads being generated as a result of their videos, allowing them to do better in the long run.

Therefore, this will give our client a better gauge of how much buzz their videos are generating, and if the buzz analysis shows a sharply declining trend, they can opt to garner more attention for their videos from various other marketing channels in order to shore up the number of comments they receive.

Survival Analysis (Facebook)

To order to extend our analysis further, we have also performed a survival analysis based on the comments retrieved. This is done using the Kaplan-Meier estimator. For the purposes of the Kaplan-Meier estimator, two crucial variables were needed - the event occurrence and the duration. For Facebook, the event occurrence that we determined was whether a post still received comments after 24 hours have elapsed since its posting. Posts that do have comments after 24 hours would have been deemed to have “survived”. The duration was the difference in the number of hours elapsed between a Facebook post and when the comment was made.

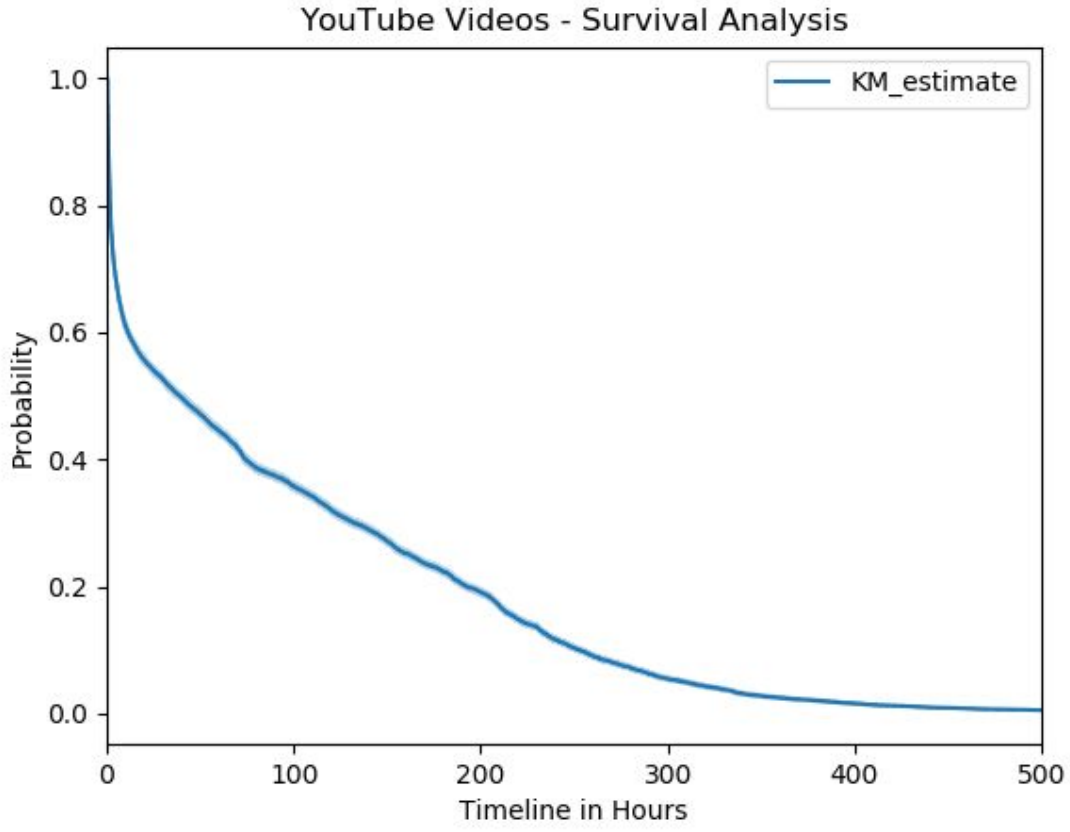


Survival Analysis for Facebook Posts using Kaplan-Meier Estimator

Therefore, the above graph shows that within the first few hours, a Facebook post has a much higher chance of “surviving”, but it quickly tapers away, with a low probability of approximately 10% chance of a post receiving comments after 10 hours, and continues decreasing gradually as the hours go by.

Survival Analysis (YouTube)

For YouTube, we also performed a survival analysis on all of our client's videos. Similarly, the event occurrence that we determined was whether a video still received comments after 24 hours have elapsed since its upload. Videos that do have comments after 24 hours would have been deemed to have "survived".

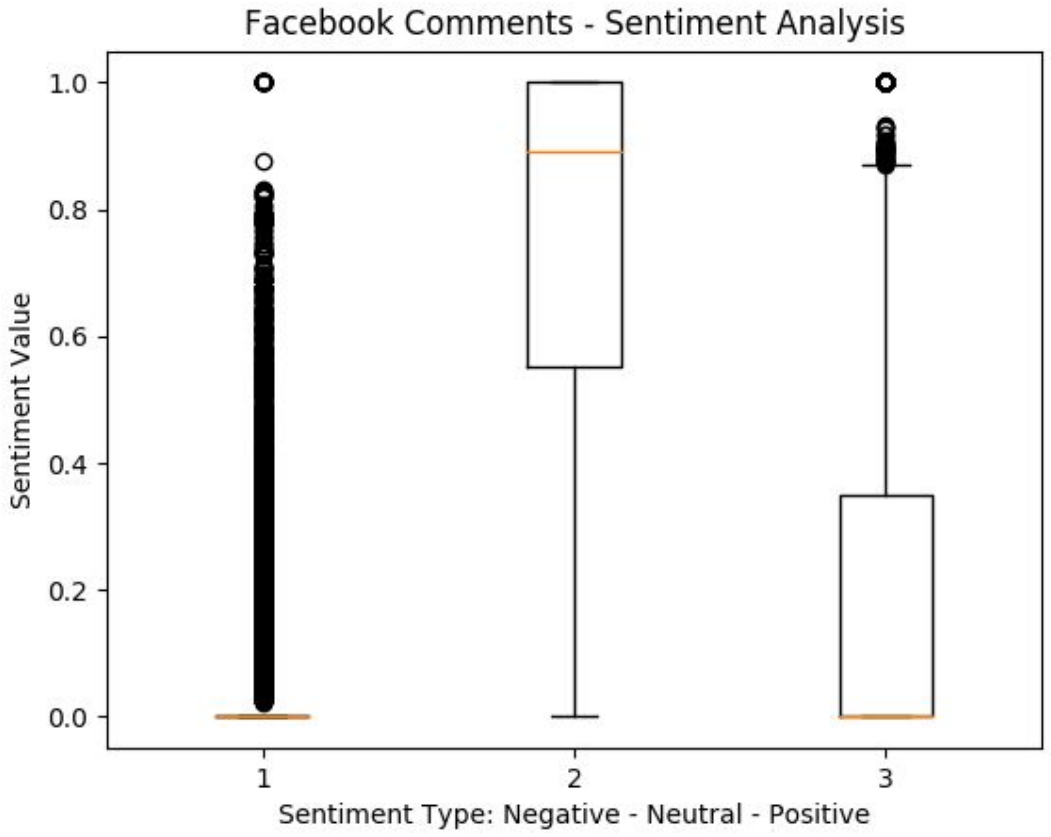


Survival Analysis for YouTube videos using Kaplan-Meier Estimator

One ostensible difference for the YouTube survival analysis is that the survival probability for YouTube videos is much higher than that of the Facebook survival analysis, in which it takes approximately 500 hours for the graph to plateau completely. Contextually, this makes sense as the bulk of our client's followers would be more active on YouTube, where the videos are uploaded, and not Facebook, and provides a better gauge for our clients regarding the probability of a video surviving a given number of hours.

Sentiment Analysis (Overall - Facebook)

Additionally, through the use of a lexicon and rule-based sentiment analysis tool known as Valence Aware Dictionary and sEntiment Reasoner (VADER), we were able to achieve much greater accuracy of sentiments and also segment them into 3 different categories - whether the comment was negative, neutral or positive. Furthermore, VADER was especially useful as a means of multigram polarity analysis, over a unigram analysis as it would be able to better capture sentiments that are expressed over short phrases as well. Therefore, this would enable Ministry of Funny to quickly have a snapshot of the audience receptiveness towards a particular video or theme.

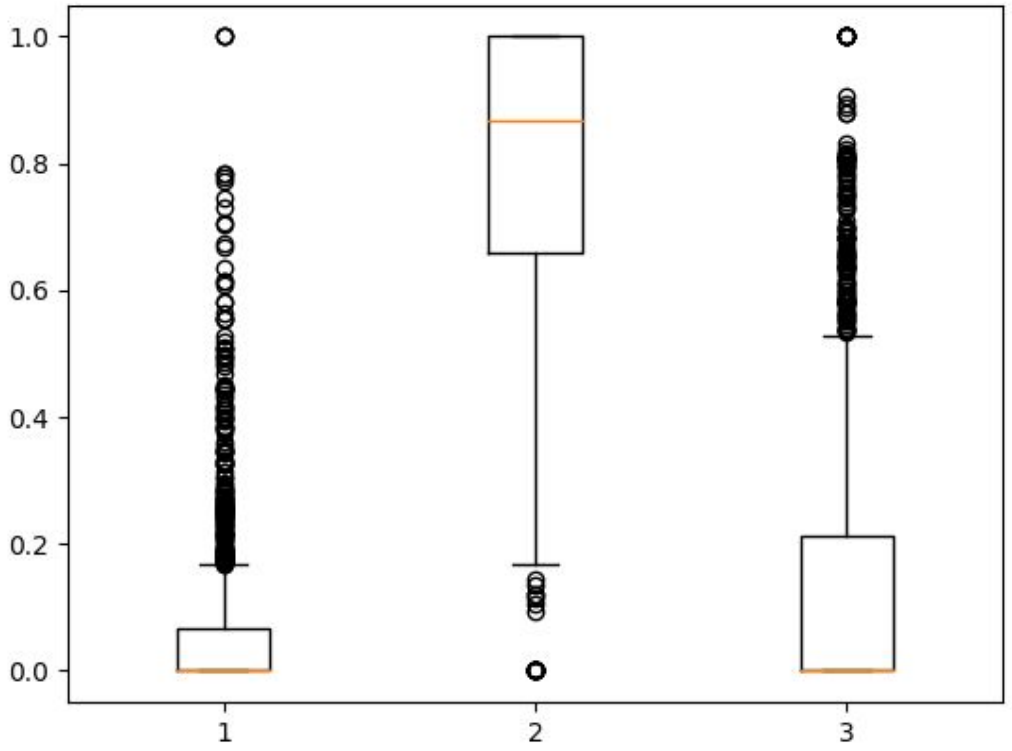


Overall Sentiment Analysis for all Facebook post comments

Overall, most of the comments are neutral in nature. However, there is a wide polarity of both positive and negative comments as well, which may allude to the controversial nature of several of their video themes. Therefore, this allows Ministry of Funny to evaluate what the general sentiment towards their videos are.

Sentiment Analysis (Individual - Facebook)

In line with the frequency and buzz analysis, we also enabled the option for our client to select a particular video or post to view its corresponding sentiment analysis.

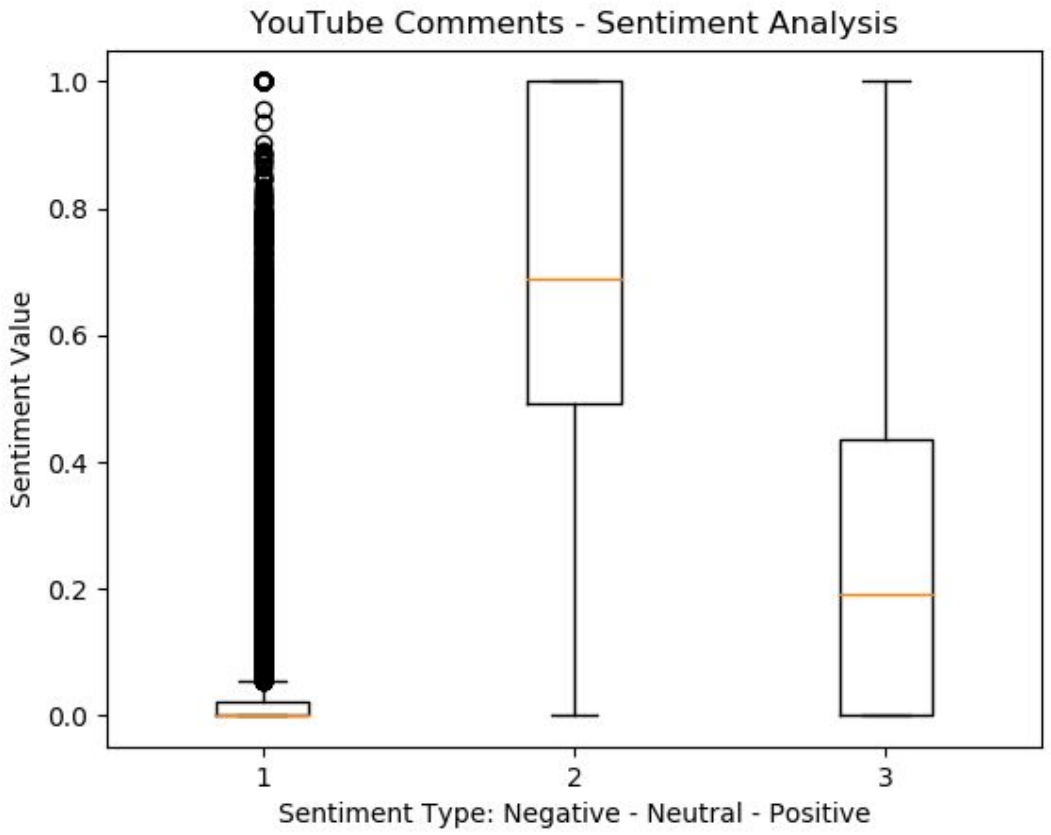


Sentiment Analysis for the Facebook Post on questions that LGBT people get

Revisiting the earlier example used for the buzz analysis, this sentiment analysis diagram shows a high polarity of comments, with a much large spread of both positive and negative sentiments when compared with the overall sentiment analysis.

Sentiment Analysis (Overall - YouTube)

In addition, we also performed a sentiment analysis for all of our client's YouTube videos as well, as demonstrated in the boxplot below.

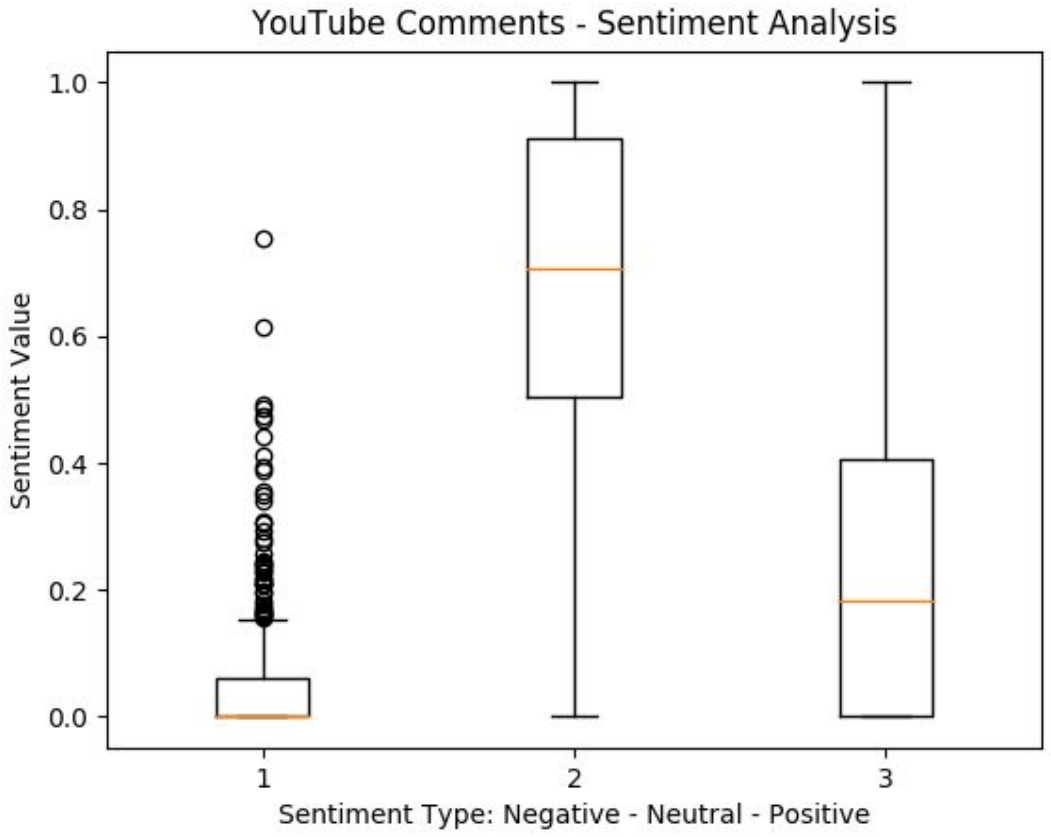


Overall Sentiment Analysis for all YouTube video comments

In comparison with the overall Facebook sentiment analysis, the YouTube analysis has more mildly positive sentiments, with a wider spread of neutral comments as well in general.

Sentiment Analysis (Individual - YouTube)

In addition, we opted to allow our client to perform sentiment analysis on a per-video basis as well. The below box plot is the sentiment analysis for the video on questions that LGBT people get.



Sentiment Analysis for the YouTube Video on questions that LGBT people get

In general, the overall sentiment specific to this video ranged from mildly negative to mildly positive, with a slightly lower spread of neutral comments when compared to the overall YouTube sentiment analysis.

Therefore, sentiment analysis would allow our client to better gauge the receptivity of their audience, and as a proxy to measure how controversial their video was, as a high polarity between positive and negative sentiments would inevitably spark more conversations which is one of their goals.

Clustering Analysis with LDA Topic Modelling (Facebook)

Additionally, we intended to further enhance our analysis with the use of LDA Topic Modelling in a clustering analysis, which would allow keywords to be assigned a certain probability to belong to a category or topic. Initially we thought that this would be especially useful with regards to Ministry of Funny's focus on having videos that spark social conversations, such as when certain buzzwords fall into a particular category or theme. However, upon running the LDA Topic Modelling analysis, we realised that we could not garner any useful information.

[[(0, '0.123*"vid" + 0.063*"share" + 0.063*"youtube" + 0.063*"ly" + 0.063*"jointheministry" + 0.063*"reference"'), (1, '0.067*"bit" + 0.067*"videos" + 0.067*"check" + 0.067*"funny" + 0.067*"360" + 0.067*"reference"'), (2, '0.067*"degree" + 0.067*"bit" + 0.067*"funny" + 0.067*"check" + 0.067*"found" + 0.067*"guys"')]]

LDA Topic Modelling for Facebook Comments

Clustering Analysis with LDA Topic Modelling (YouTube)

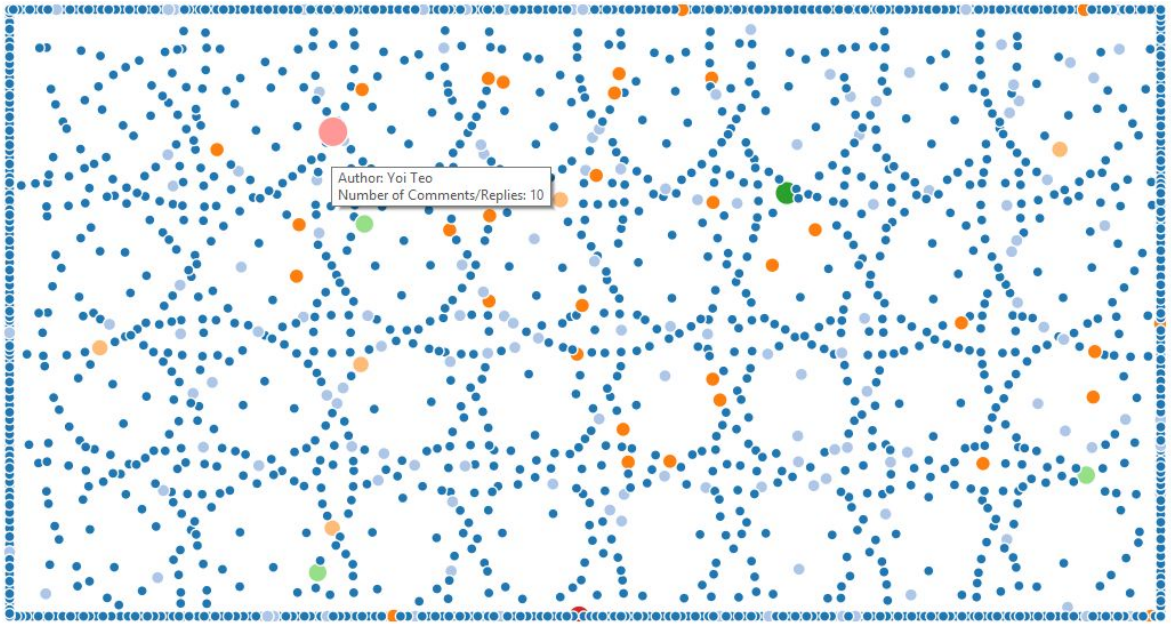
Similarly, performing a clustering analysis with LDA Topic Modelling on the YouTube comments did not lead to any meaningful results as well. In addition, the Unicode incompatibility due to emojis not being factored in could have affected the LDA Topic Modelling as well. Therefore, this may indicate that at this point in time, LDA Topic Modelling is generally not useful for them, unless in the future, they choose to revisit past themes with new videos, which will then provide a better basis for topic modelling to work out.

[(0, u'0.167*"guy" + 0.167*"looks" + 0.167*"like" + 0.167*"vice" + 0.167*"schools" + 0.167*"principal"'), (1, u'0.167*"guy" + 0.167*"looks" + 0.167*"like" + 0.167*"schools" + 0.167*"vice" + 0.167*"principal"'), (2, u'0.167*"principal" + 0.167*"schools" + 0.167*"vice" + 0.167*"like" + 0.167*"looks" + 0.167*"guy"')]

LDA Topic Modelling for YouTube Video Comments

Social Network Analysis (Individual - Facebook)

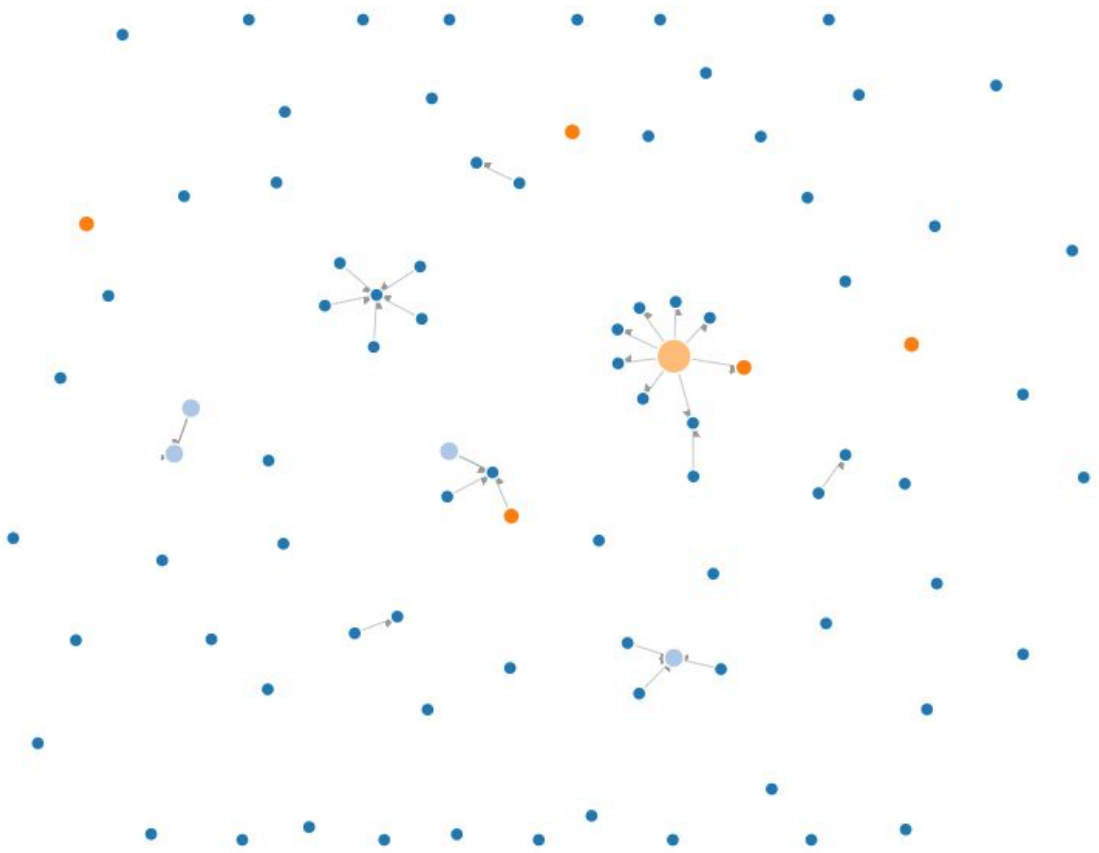
Social network analysis would enable Ministry of Funny to pinpoint key influencers whom they wish to target or reach out to, as these influencers are often able to generate a network effect, which would help to further the virality of their videos on the whole.



Social Network Visualisation of Facebook Comments on free transport video

Through the visualisation above, the Facebook network is sparse, and made out of isolates. In order to help our client identify key influencers, we have also enabled a tooltip and the different colour/size of the node will help to differentiate it as a key influencer.

Social Network Analysis (Individual - YouTube)



Social Network Visualisation of YouTube Video Comments on free transport video

For the visualisation of YouTube comments, even though the network is more sparse due to the lower number of comments for this particular video, there are many different network structures such as dyads and stars. Therefore, our client can more easily identify buzz generators in which a large number of people reply to them, and can reach out to such individuals more easily.

Dashboard

In order to provide more business value to the client, we have opted to present the visualisations for the various analyses above on a dashboard, segmented by Facebook and YouTube. Additionally, as mentioned above, our client will be able to choose which post or video to visualise, in order to have a more granular view as well. Once a post or video has been selected, the corresponding visualisations will be generated for the selection. A simple login and logout function will also be implemented for data privacy and security reasons. In order to ensure quick processing speeds, the dashboard will also be hosted on an AWS server located in Singapore.

Conclusion

In the long run, we believe that this dashboard will give them a powerful means of analysing their performance, and allow them to gleam data driven insights from these analyses.

Team Contributions

The following table illustrates the work distribution across our team members for the midterm milestone:

Team Member Name	**Tasks Done**
Vanessa Goh Ze Hui	1) YouTube Analyses & Data Collection (inclusive of Facebook Social Network Analysis) 2) Setting up Database & ER Diagram 3) Final Report & Slides 4) Dashboard & Visualisations
Wu Yunheng (Winston)	1) Facebook Analyses (exclusive of Social Network Analysis) & Data Collection 2) Final Report & Slides 3) Client Liaison