INSY 670 GROUP PROJECT

YINGXIN JIANG, KEJIA LIU, KEXIN WANG, YICHEN WANG, DIWEI ZHU

I. Introduction

In this project, we will explore what role the social media (Twitter) played during the famous Ice Bucket Challenge campaign in 2014 and how it influenced the trend of this charity movement from two perspectives: first, we will apply methods such as topic modelling and sentiment analysis to investigate the thoughts and attitudes of Twitter users had towards this campaign. By integrating scores generated with time labels, we could come up with detailed trend plots to gain a deeper understanding. Second, network analysis will be used to examine how the ALS association and other influencers/ celebrities contributed to this campaign.

II. Background

As one of the incurable and fatal diseases, Amyotrophic lateral sclerosis (ALS) is currently affecting more than 200,000 people around the world, where patients will gradually become paralyzed due to loss of communication between the brain and muscles. This deadly disease had received little attention around the globe until 2014 when an unprecedented campaign called Ice Bucket Challenge was raised on the Internet. Participants would pour cold water over their heads to experience the “freeze” symptom of ALS and challenge their friends to do so as well to raise attention from society. The campaign has gone viral in only several months and become one of the most influential charity movements in internet history.

III. Social Media Data Acquisition and Pre-processing

Data acquisition

To obtain the required dataset, we perform data scraping over Twitter, both on related tweets and user info of all mentioned users using Twitter Developer API.

For tweet scraping, all the tweets mentioning #IceBucketChallenge are collected, where data ranges from July 2014 to December 2014 since this campaign achieved its popularity peak during this period according to the tweet summary statistics. To reduce the data scraping time cost, we only chose the 1st, 15th and 30th of each month to represent important milestones to reduce dataset size. And there's a limit of 5000 tweets per day due to the possibility of having an enormous amount of related tweets posted on that day. These tweets will be used for sentiment analysis, topic modelling as well as network analysis.

For all the users who posted tweets or got mentioned, we have scraped their user info as well to come up with a dataset including the number of followers, listed count and language used. These data could be used as equation variables for determining top influencers. After scraping we did some data preprocessing including adding dummy variables on hashtags and whether certain topics are mentioned.

IV. Topic and Sentiment Analysis

Campaign popularity and sentiment trend analysis

To reveal the popularity change of the campaign on Twitter over time, we first extract the dates of each tweet we collected and then plot the number of postings to visualize the trend (Figure 1).

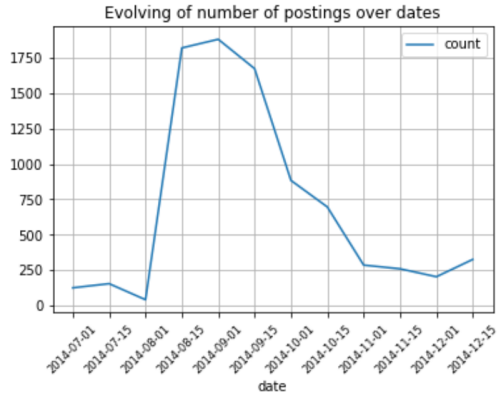


Figure 1. Popularity change of the ice bucket challenge campaign

The boom in popularity happened in the first half of August 2014 and reached its peak at the start of September. Then, the number of discussions on Twitter gradually and exponentially decreased as time went by. At the end of the year, the popularity went higher to a small extent, referring to users' end-of-year wrap-ups or thaw-backs as possible reasons.

Furthermore, as allowed by hashtag topics – the special feature of Twitter that undoubtedly states the social media characteristics, we extracted the top 5 popular hashtag topics from the tokenized texts of the scraped tweets: #icebucketchallenge, #als, #alsicebucketchallenge, #strikeoutals, and #mnd. A graph on the evolution of each hashtag with respect to time is shown here below (Figure 2). We see that #icebucketchallenge has comparatively high popularity, with #icebucketchallenge embodying a similar trend as the tweets

posts' trend obtained from the previous step. The popularity of #als also increases at the beginning of our observed period but dropped significantly after a month.

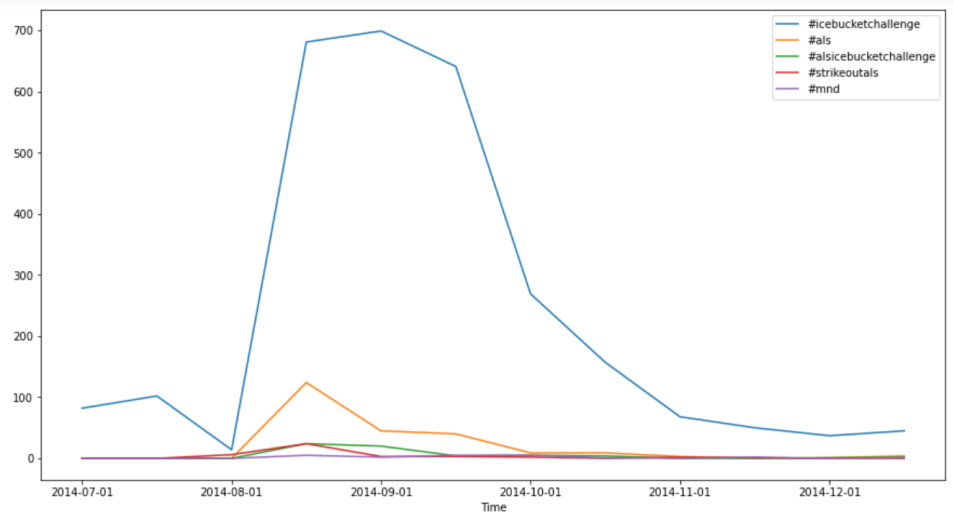


Figure 2. Popularity change of top 5 hashtags

With the *vader* package, we calculate the compound sentiment scores for each of the tweets. We classified each tweet's sentiment into these three categories: positive, neutral and negative. Then, we plot out a line graph denoting the change of sentiment over time. We see that the peak of positive tweets appears first, followed by the peak of neutral tweets. The number of negative tweets stayed comparatively stable from August to October and experienced a peak during mid-October. Moreover, the negative sentiment even rises at the end of the year. Overall, this reveals a decreasing satisfaction and support for the ice bucket challenge. However, the number of positive posts still exceeds posts with neutral or negative sentiment.

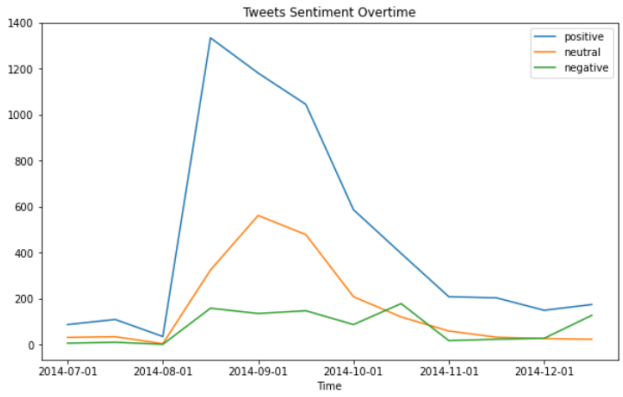


Figure 3. Tweets Sentiment Overtime

Topic modelling

To further extract the hidden structure behind this collection of tweets, we used the Latent Dirichlet Allocation model (referred to as LDA). Given the number of documents and number of words, we determined the number of topics and ran the model. The output includes both the distribution of words for each topic and the distribution of topics for each document. In our case, each document refers to a tweet post. Utilizing Python packages such as gensim, nltk, pprint, pyLDAvis and pickle, as well as a for loop on the number of topics, we were able to find the optimal number of topics and generate a dashboard for our topic modelling result (Figure 4.). We identified 3 topics in total. Among all topics, the top words used are Youtube, http and als. However, there still exists a slight difference, as topic1 focuses on the challenge itself, topic 2 are non-English tweets and topic 3 focuses on likes and tagging.

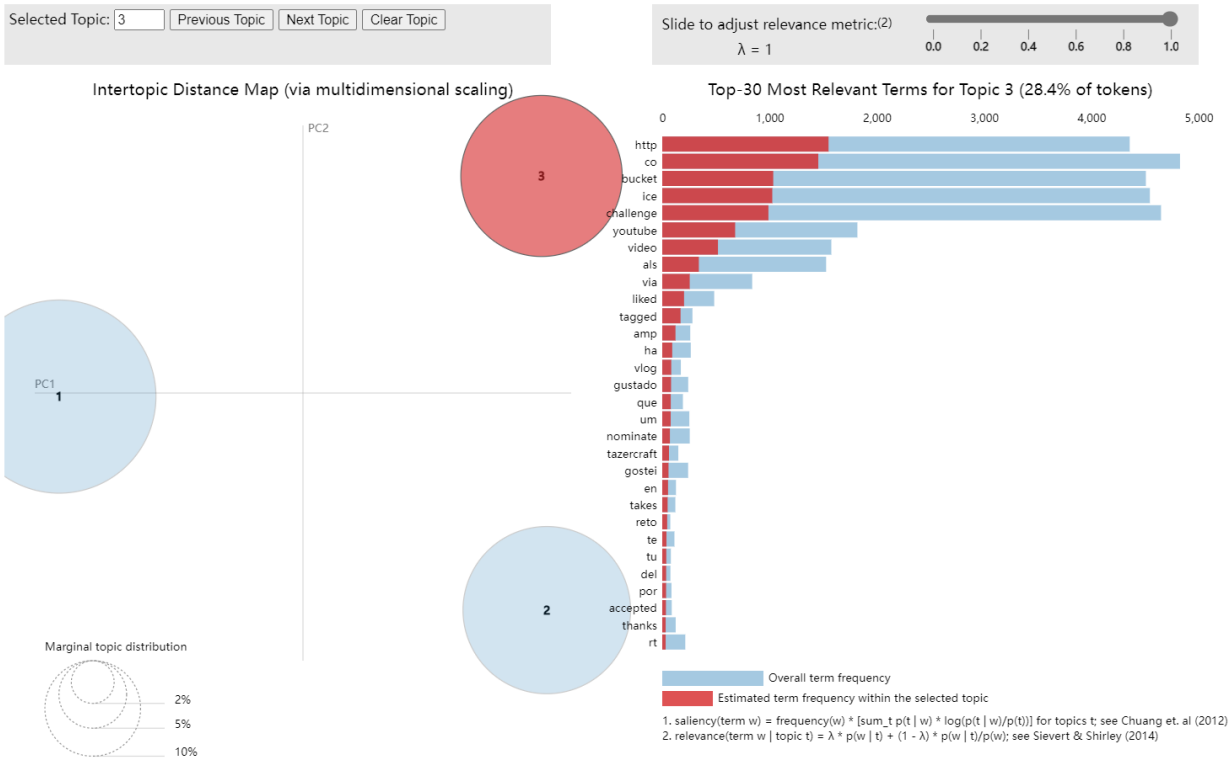


Figure 4. Topic Modeling Result

Subtopic analysis

To undertake a more in-depth topic analysis, we identified two subtopics with keyword dictionaries generated by us. The ALS subtopic, linking to keywords like *muscle*, *disease*, and *patient*, labels tweets that paid attention to the ALS disease. A mention of keywords like *donate*, *dollar*, and "*$*", makes the tweet be labelled under the charity subtopic. Among the 8016 tweets we collected, there were only 91 (1.1%), 230 (2.9%) and 19 (0.2%) tweets that were under the ALS, charity, and both ALS and charity subtopics respectively. 95% of the tweets did not mention any of the subtopics.

Integrating with the sentiment analysis result: we compare the average sentiment scores of the tweets under different subtopics (Figure 5). The average sentiment scores of the

ALS subtopic and non-topic are 0.197 and 0.204 respectively; the average sentiment scores of the charity subtopic and ALS&charity subtopic are 0.287 and 0.271 respectively. The result from the follow-up two-sided T-test proved that the charity subtopic indeed relates to a higher average sentiment score (p-value = 0.000112), while the link between the ALS subtopic and lower sentiment scores is not statistically significant (p-value = 0.618137).

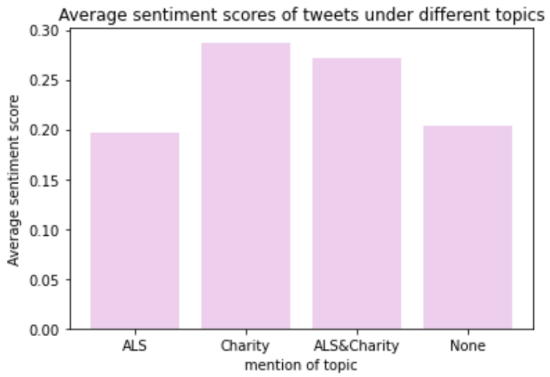


Figure 5. Average sentiment scores of tweets under subtopics

Integrating with numbers of mentioned users: were the tweets under the subtopics easier or more difficult to spread among users caught our attention. We generate the average number of users mentioned per tweet for each subtopic category (Figure 6). The average numbers of mentioned users per tweet are 1.26, 1.49, 1.32, and 1.53 for tweets that were under ALS subtopic, under charity subtopic, under ALS&charity subtopic, and link to no subtopic. The two-sided T-test shows that there is a statistically significant relationship between a lower number of mentioned users and ALS subtopics (p-value = 0.000004), while fails to verify that tweets under the charity subtopic significantly mention more users compared to other tweets (p-value = 0.410627). The result may indicate that tweets under ALS subtopics were more isolated and provoked fewer user interactions.

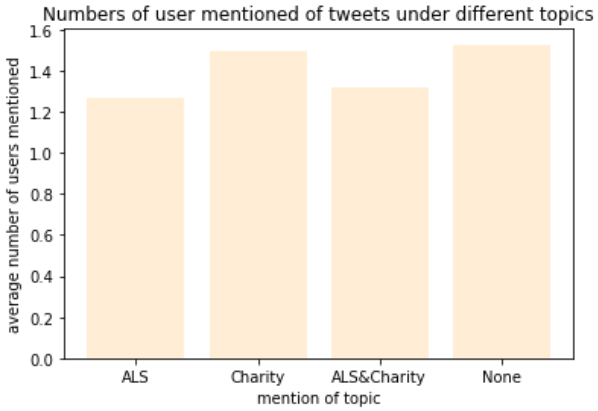


Figure 6. average numbers of users mentioned per tweet under subtopics

V. Network Analysis

We scraped 5,000 Tweets with Keyword=" Ice Bucket Challenge" and their corresponding username from July to December. After the data preprocessing, we counted tweets and retweets based on the scraped data, so we generated the users' interaction dataset. Now we can take a look at closeness, degree and betweenness centrality and realize that most of the results are 0 which is not helpful for further analysis. Therefore, for gaining more useful features, we scrap user information including list_counts, followers_counts, language, etc. Since so many values on language are missing, we decide to only keep list_counts, and followers_counts. After that, we turn to count how many times a user is mentioned by others, and for all users that posted or were mentioned. Then, we combine the number of mentions as a new feature with the previous features together, and make a standardization on it, so that one of the features wouldn't be domain on the further calculation.

We make a reference from previous predictive feature importance of the model and use our own judgment, we pick w1 = 0.182173, w2 = 0.111455 and w3 = 0.114590 for the number of

lists, followers and count mentions. Then, we balanced the weights with ratio (making sure that w_list+w_follow+w_mention=1) and subjective ±0.5 adjustment.

$$ratio = 1 / (w1+w2+w3)$$

$$w\_list = w1 \times ratio + adjust$$

$$w\_follow = w2 \times ratio$$

$$w\_mention = w3 \times ratio - adjust$$

With the assigned weights and the normalized user information, we identified the top 100 influencers of the keyword with the scores calculated through:

$$Score = w\_list \times listed\_count + w\_follow \times \#followers + w\_mention \times mentions,$$

where w_list+w_follow +w_mention = 1.

By interpreting our analysis results, we can generate many meaningful insights. Our analysis indicates that the top 10 influencers of this challenge on Twitter are: Justine Bieber, YouTube, Barack Obama, Lady Gaga, NYTimes, cnnbrk, Katy Perry, Taylor Swift, CNN, and Rihanna. Justine Bieber seems to be the most influential figure in this challenge, possibly since he was one of the most popular singers with a huge fan base back in 2014. Obama was also a top influencer in this challenge, the challenge was so well-known that it even brought the attention of the then-president of the United States. In addition to celebrities and political figures, media such as YouTube, New York Times, and CNN also play important roles in spreading the influence. However, none of the ALS authority Twitter accounts are considered top influencers according to our analysis.

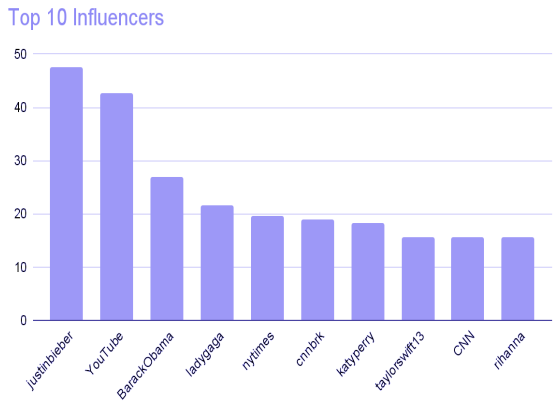
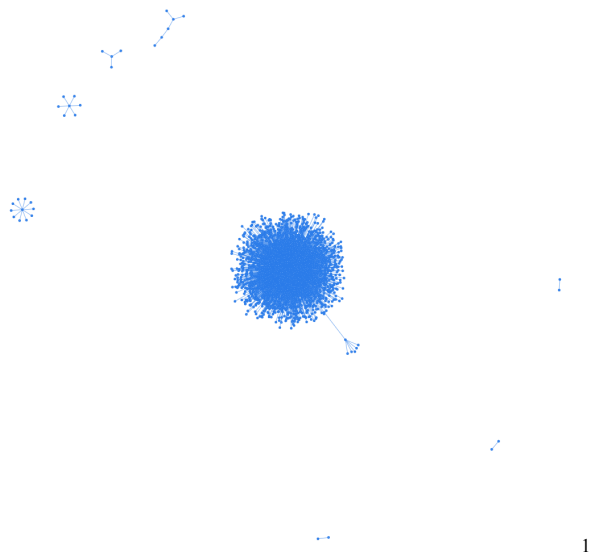


Figure.7 Top 10 Influencers with statistical count

To better understand the results, we created a visualization for networks of the top 10 influencers that we identified in the ice bucket challenge, using the Pyvis library. According to the network visualization shown below, the big circular cluster is the network of youtube. As a popular media platform, YouTube was mentioned the most when users posted tweets regarding the ice bucket challenge, and thus had the largest network. This is reasonable because many people are either watching or posting ice bucket challenge videos on YouTube. Other smaller surrounding clusters are the network of Justine Bieber, Katy Perry, Taylor Swift, Lady Gaga, and Obama. Unexpectedly, these top influencers showed smaller networks. This is probably because there are not a lot of postings that mention them in the dataset we scraped. However, the size of the network does not necessarily represent how powerful the network was. For instance, The Laugh Factory, which is the official Twitter account of a popular TV show, is part of Justine Bieber's network; and Rihanna, the famous singer, is part of Lady Gaga's network. Although the network visualizations of the celebrities and political figures appear to be small, it does not mean that they are not powerful or influential.



[1]

Figure.8 Network Structure

Overall, we observed that the engagements of celebrities and political figures may have more impact in terms of raising awareness, and media platforms play an important role in generating larger spreads. Therefore, it might be a good idea for ALS organizations to collaborate with media or celebrities to gain more attention, educating people more about the ALS disease rather than just the challenge itself. Other health organizations and non-profit organizations can also consider this approach to raise awareness.

[1] The network visualization generated using pyvis is dynamic, please see the html file for the complete visualization gif and more details if needed.

VI. Recommendations and Insights

Recommendations

Based on the results and findings above, we proposed a strategy for campaigns with social media platforms similar to Twitter (Figure 9.).

Figure 9. Proposed Strategy for Campaign with Social Media

The full strategy is divided into two phases: initiation and management. Each phase can be broken down into two stages. Each stage can be separated into two detailed steps. For phase 1, the first stage is to create an initial design for the campaign, as well as set up accounts. At this stage, campaign organizers should define the linkage between the campaign objective and tools for escalation. Here, entertainment-related tools are especially fast when it comes to wide-spreading information, which is the case with the ice bucket challenge. The campaign organizers should also get popular social media accounts involved, which includes sending invitations to influencers to help with promotion. The second stage of phase 1 is the first round of testing of the initial design. Data on social media platforms' reactions will be collected, and an

initial analysis will be performed. Based on the analysis results, the design should be revised for scaling up.

Next, we enter the second phase, which includes stage 1's overseeing and controlling of posts, as well as stage 2's continuation of popularity. For stage 1, the campaign organizers should frequently check on hashtags and popular accounts' performance to ensure the effectiveness and minimum deviation from the initial objective. This step can be done by regularly scraping tweets and performing tests on sentiment analysis, topic modelling, etc. Based on the results from the analysis, organizers can determine and properly allocate an amount of effort to achieve optimal campaign outcomes. Stage 2 begins when a decline is observed. By methods such as adding new content and involving new influencers, organizers can manage the slowdown of the campaign gradually instead of a sharp exponential decrease. Organizers can also create a campaign "wrap-up" to fully utilize the year-end social media boom.

By adopting this strategy, campaign organizers can expect growth from three perspectives: effect lifting, including popularity/revenue uplift, response rate uplift, social media engagement uplift; cost-saving with a wider target range; operational efficiency with reduced possibility of deviation from the campaign's main objective.

Insights

The results from the topic analysis reveal that the UGC discussions under the ALS disease and charity subtopics were very rare. In the visualized network, we found no ALS-related associations or authorities, nor scientists or doctors. Rather, it was the celebrities and political figures who attracted most of the attention.

The form of the challenge – find a bucket of ice, pour the ice over your head, post the video online, and wait for funny comments and forwards – is so entertaining and easy to replicate, and the challenge did actually rely on celebrities to expand its influence. The result pushes us to ponder whether the users participated in the challenge just for entertainment, and, as argued by the famous book *Amusing Ourselves to Death*, whether the form of the challenge diluted or excluded the actual goal of itself. Although the original book built its argument based on an analysis of television programs, recent voices state that there is a "transition" or "swap" from television to social media while having the "amusing ourselves to death" phenomenon unaltered. We start to question if our project becomes an attestation of this critique on social media and proved the failure of the ice bucket challenge on Twitter.

To our ease, according to the ALS Association, 150 million donations from people and 90 million research funds given by governments were committed after the boom of the challenge. Accordingly, the association funded research, made the peer-viewed publications increase by 20%, and built one of the largest resources of ALS whole genome-sequencing data that has been shared with partners all over the world[2]. The accelerated pace of fighting against ALS, as thrillingly announced by the association, has an undoubted link to the challenge.

We refer to several facts that can shed a light on the disparity between our results and the actual condition. Firstly, people's awareness can be formed and actions can be encouraged through multiple channels, not just social media. Second, the awareness of the ALS disease is difficult to quantify and is not always reflected on social media. Therefore, considering that we rely only on a small dataset collected from just Twitter, it is understandable that our result cannot reflect what the ice bucket challenge actually provoked.

[2] https://www.als.org/stories-news/ice-bucket-challenge-dramatically-accelerated-fight-against-als