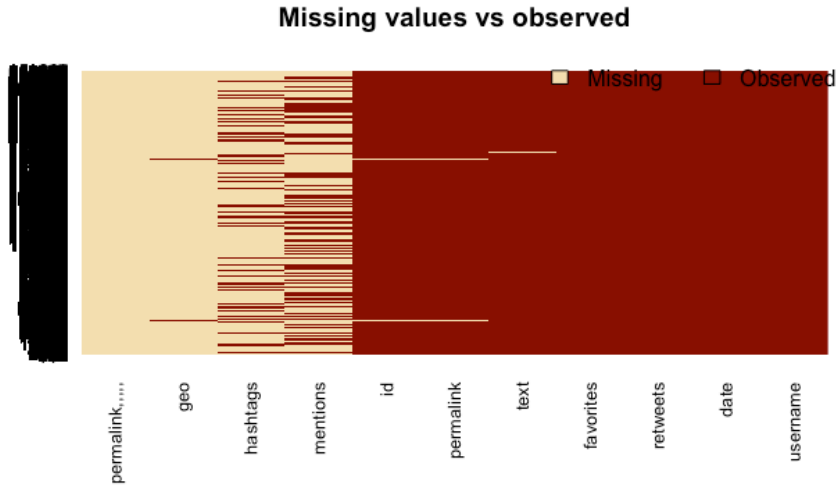
Method & findings

Sample

This study uses the programming language python to scrape all of the tweets within two weeks of time (08/21 – 09/04) with the crisis escalation date (08/28) in the middle. The python package this study uses is called “GetOldTweets-python.” The twitter API limits the users access by providing maximum 3200 related tweets with the oldest them no earlier than a week ago. This python package can bypass the limit of Twitter API to fully assess the whole dataset of related tweets. The query of the tweets are searched by key word “Joel Osteen” because the leading pastor Joel Osteen was the main attributor of the blame suggested by the pilot study of the tweets and mainstream media coverage. The researcher downloaded all of the tweets from 10/20 – 10/27. In total, there are 259632 tweets. The sampled tweets are then cleaned and manipulated in R. The resulting dataset is a 259632 * 17 data frame. This dataset can be visually viewed in the following picture. The x-axis displays the key variables. And the corresponding color represents if the data is missed. As the picture suggests, the key variables such as texts, favorites and retweets are clean and complete while unfortunately, geo information of the users are not documented.



“Bag of word” text mining

The bag-of-words model is a simplifying representation used in natural language processing (NLP) and information retrieval. In this model, a text (such as a sentence or a document) is represented as the bag (multiset) of its words, disregarding grammar and even word order but keeping multiplicity. The bag-of-words model is commonly used in methods of document classification where the occurrence of each is used as a feature for training a classifier.

Based on this model, the researcher is able to perform certain statistical methods on the text data.

Naïve Bayes classification algorithm

A Naive Bayes classifier works by figuring out the probability of different attributes of the data being associated with a certain class. This is based on bayes’ theorem. Naïve Bayes has been proved to be a classic and reliable way to conduct sentiment statistical learning on the text data.

Result

Distribution of the tweets

The distribution of the tweets is as the following figure 1. This research will review some key responses by Joel Osteen when explaining the distribution of the tweets.

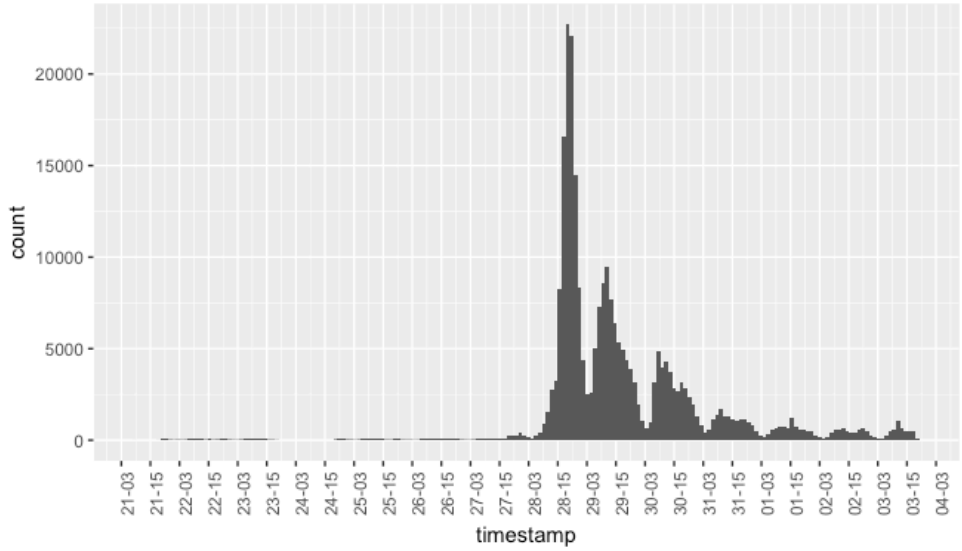


Figure 1. Distribution of tweets in time-series

As figure 1 suggests, the twitter explodes on the 28th from almost 9pm. In total, there were 98878 tweets, more than a third of the torrents of all 259632 tweets generated in the two-week window time. The very start would be back to 08/26th, when Joel posted on his twitter at 1:30pm that says:

"Victoria & I are praying for everyone affected by Hurricane Harvey. Please join us as we pray for the safety of our Texas friends & family."

This statement drew some criticism immediately in a small scale. Some tweets state that prayer is not useful during this situation and Joel Osteen should do more than simply offering prayers. This is still at the pre-crisis stage in which most tweets do not carry much influence weights among all of the tweets generated in the two weeks. A correlated topic during this period was showing the wealth of Joel Osteen especially his luxurious mansion to suggest his inactivity. Some more related criticism came later but still in a small scale. On 08/27th, the church posted:

"Dear Houstonians! Lakewood Church is inaccessible due to severe flooding! We want to help make sure you are safe. Please see the list below for safe shelters around our city, and please share this with those in need!"

This post initiates a heated discussion about whether the church has been flooded. And since then the sheer amount and the negative valence of tweets escalated in an unprecedented way.

On the 28th, Joel released a second statement through the Lakewood website:

"We have never closed our doors. We will continue to be a distribution center for those in need. We are prepared to house people once shelters reach capacity. Lakewood will be a value to the community in the aftermath of this storm in helping our fellow citizens rebuild their lives."

On the 29th, he tweeted:

"Victoria and I care deeply about our fellow Houstonians. Lakewood's doors are open and we are receiving anyone who needs shelter."

On the same day, an influential user (will show the figure of influential users below) who criticized Joel at the early stage posted a photo about lines of air mattress in the Lakewood church, saying that the church was "prepping to open its doors".

On the 29th, the church confirmed on Twitter that it was open to evacuees and taking in supplies.

From the 29th, tweets related to Osteen plummeted even though it was still beyond the normal posts. Then the tweets declined fast to the almost same level as before the crisis happen.

Distribution of sentiments in time series

Using Naïve Bayes learning model, this research predicted the sentiments of all tweets. The predictive model arrives at 72.5% accuracy rate with the 95% confidence interval (61.38%, 81.9%) (Note: this model reaches pretty good sensitivity in predicting the negative sentiments but suffers greatly in predicting the positive sentiments, resulting in an underreporting of the

positive sentiments). The distribution of positive vs. negative sentiments is in the following picture:

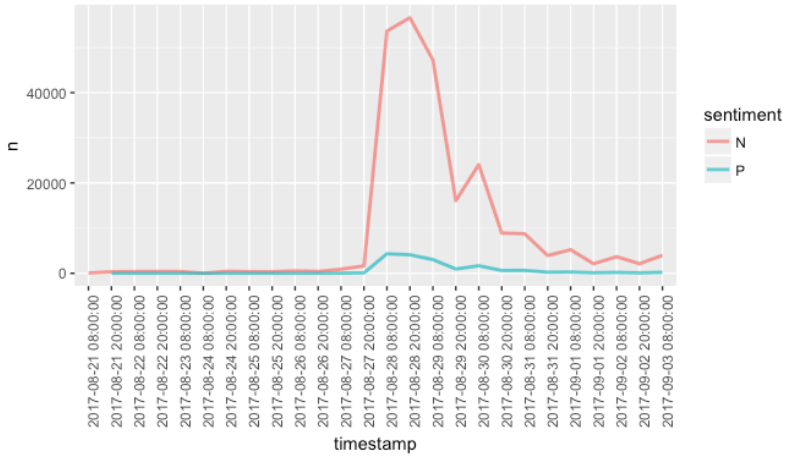


Figure 2. Distribution of sentiments in time series

In the figure 2, the red curve denotes the negative sentiments while the blue curve represents the negative sentiments. In this crisis context, any tweets involving questioning of Joel's motive for not opening the church, urging him to open the church or even questioning of his spirituality and capability as a pastor are categorized as negative sentiments. Tweets involving praising his spirituality, his good deeds and defending him for not opening the church are categorized as positive sentiments. The graph shows that as approaching the 28th, the negative sentiments escalated greatly. At the same time, there is a slight raise of the positive sentiments as well. This positive raise is expected: while majority of the tweets were challenging Joel Osteen, a faith-holder group was also defending Joel especially by arguing that the church was flooded as well.

Influential users in time-series

To pinpoint the influential users, this study uses a simple formula involving two most variables collected along with the text data of tweets: retweets and favorites. Because in practice, the number of favorites is almost two times of retweets. Therefore, this study doubles the weights of retweets variable resulting the influence index being calculated as:

Influence = favorites + retweets * 2. With this index, the researcher plotted the 50 most influential (within the sampled two-week window time) tweeter usernames in time-series as it is in the figure 3.

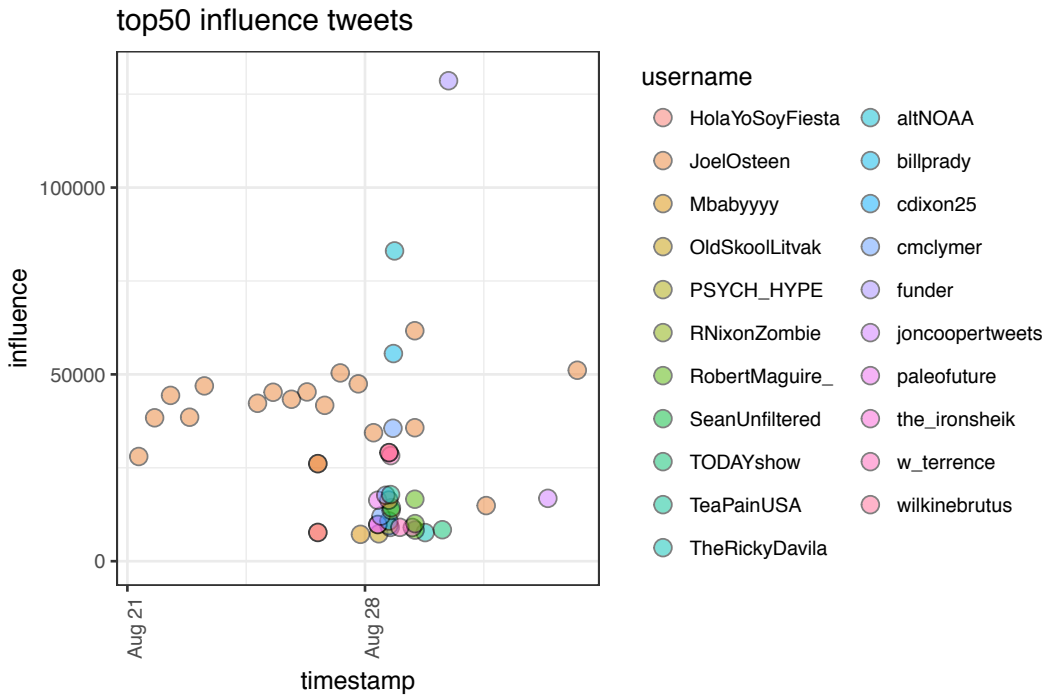


Figure 3. 50 most influential usernames in time-series

As the figure suggests, almost before the 28th when the paracrisis escalated, Joel Osteen's tweets was the only one that carries great influences and they remain relatively stable around 50,000. Joel tweets some inspiring gospel related quotes every day and this is the source of his daily influence on the figure 3. As approaching the 28th, there was another dot appear in the figure representing another twitter user. In fact, this was one of the criticism at the early stage on the 26th when Joel just posted to offer prayers. On the 28th, as the tweets increase exponentially, more influential users emerge. The most influential tweet has the influence over 120,000. And it fact that this user has 258k followers on twitter. As the number of tweets decline after the 28th, so do the influential users. And Joel's influence climb back and at the end of the window time, there was one that recovers back to the 50,000 line.

Figure 4 shows the distribution of influence of all users in time-series. This result coincides with distribution of the number of tweets, and because of the negative sentiments dominates most of the tweets, the distribution of negative sentiments.

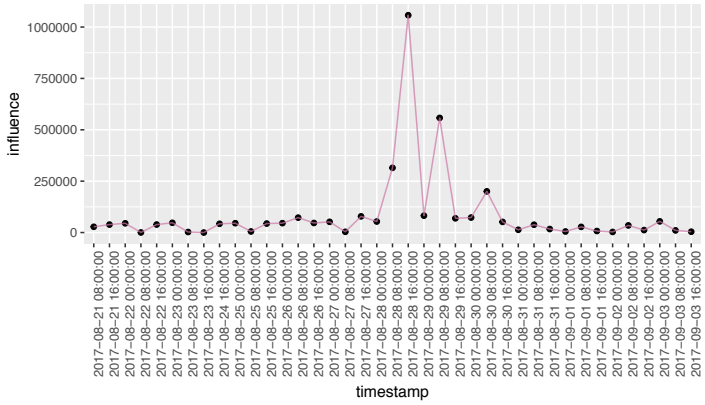


Figure 4. Distribution of influence of all users in time-series

Tweets content

To get a representative understanding of the tweets, the researcher first systematically sampled and read through 1000 tweets from all the tweets. This provides the researcher enough background understanding of the data which are important for further inference of the statistical analysis of the texts. Using the "bag of words" text mining method, this research reveals the data in the following figures.

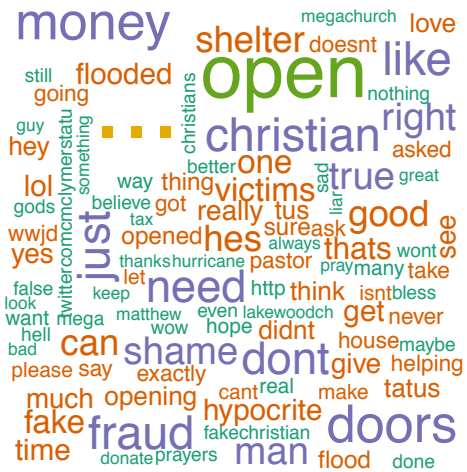


Figure 5. Word cloud for the whole dataset

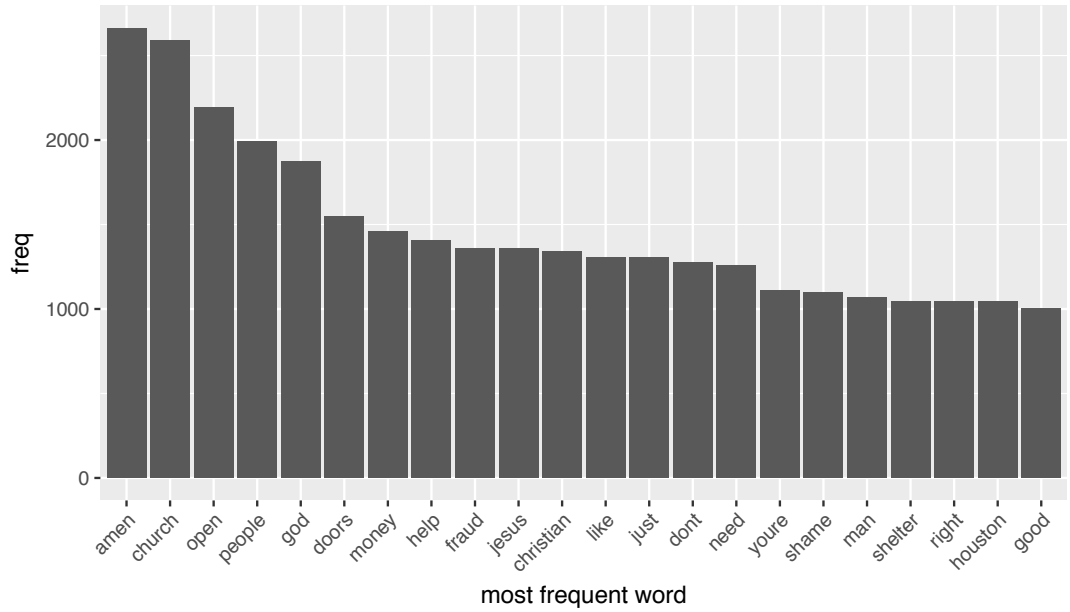


Figure 6. Most frequent words

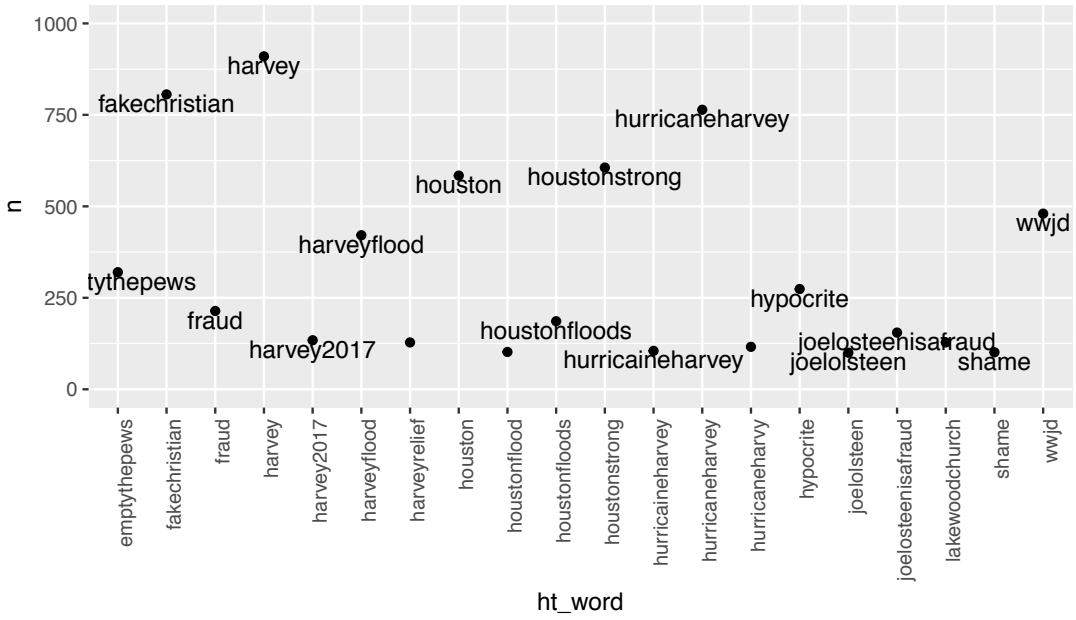


Figure 6. Most frequent hashtags

Based on the understandings of the tweets contents, this study extrapolates that the twitter contents can be roughly categorized into three groups. The first is the negative criticism of Joel for not opening the door. This is reflected by the key words such as “fraud,” “shame,” “open” and “doors.” This group can be considered mainly from the hate-holders of

Joel and the Lakewood church. The second group is the supportive or agreeing comments before or during the crisis. This is reflected by the key words such as “Amen,” “God,” “Jesus,” and “right.” This group can be considered mainly from the faith-holders of Joel and the Lakewood church. The final group is the debate over whether the door was open or whether Joel made mistakes. This can be seen from the key words such as “open,” “doors,” “help,” and “need.” This study acknowledge that this is just a rough guess about the prevalent topics among the tweets and those words do not exclusively belong to one certain group. For example, the words “open” and “doors” can be used in all three groups. For hate-holders, these words can be the reason they are dissatisfied with Joel and therefore vent their criticism; for the faith-holders, these words can be used to justify that as Joel claimed, the door was always open and Joel was prepared to help.

To further make sense of the meaning of the tweets, this study conducted cluster analysis from the document-term matrix of all a random sample of 2000 tweets (the reason not using all of the data was because my Mac does not support processing that amount of data and R would froze and crush over huge amount of data). The results can be viewed in the figure 8 and figure 9.

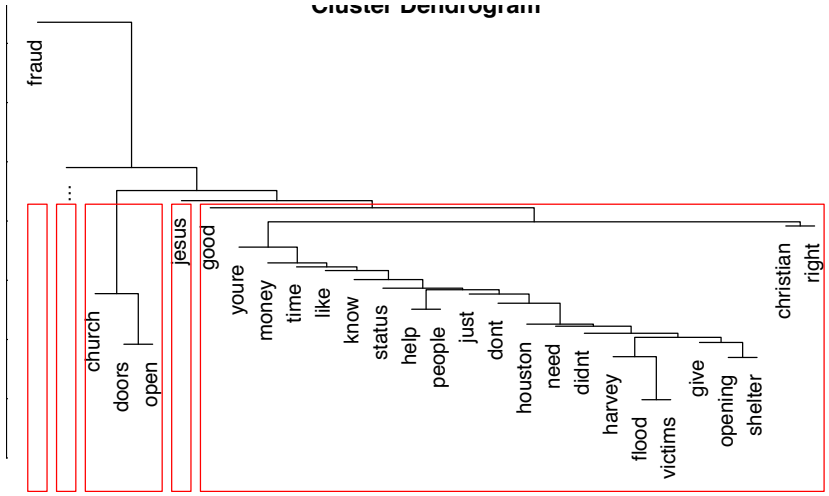


Figure 8. Cluster Dendrogram

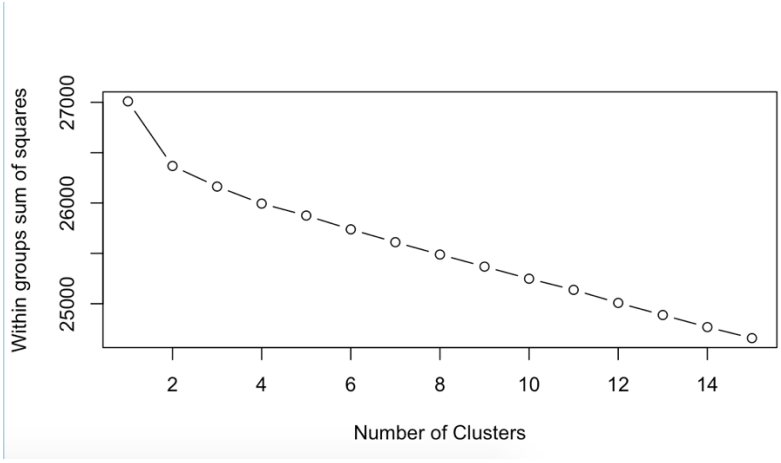


Figure 9. The scree plot for the number of clusters

From the results shown in the figure 8 and figure 9, this research makes a slight more robust claim that there are three clusters/groups of topics (because the computer does not support cluster analysis in huge volume of data, this small sample number may not be representative. Further sampling and model tweaking is needed). Following this line, this research decides to conduct a k-means cluster analysis of the potential topics. Figure 10 shows the results of those cluster results.

Not surprisingly, these results do not differ much from our rough extrapolation from the word cloud and frequent words graph. The first cluster seems to defending Joel Osteen and show supportive comments. The second cluster discusses the openness of the church and the corresponding help provided. The third cluster are mainly questioning and criticism of Joel's spirituality and his virtue and morals.

```
>
> for (i in 1:k) {
+   cat(i)
+   cat(paste("cluster ", i, ": ", sep = ""))
+   s <- sort(kmeansResult$centers[i, ], decreasing = T)
+   cat(names(s)[1:8], "\n")
+ }
1cluster 1: amen hallelujah actually helping like lot people seems
2cluster 2: spot actually helping like lot people seems thousands
3cluster 3: … church open people fraud jesus christian help
>
```

Figure 10. K-means cluster analysis (k = 3)