Topic Modeling Facebook Page Data

Anna Xu, asx2@cornell.edu

Abstract

This study explores textual data from the Facebook pages of "Business Insider," and "Humans of New York." Using MALLET Topic Modeling techniques, the study breaks the social media content into topics and analyzes which topics are popular. The findings show that the original length of the document contributes heavily to the quality of the results. It also shows evidence of certain topics receiving more engagement than others.

1 Introduction

In the era of Facebook, Instagram, Twitter, and Snapchat, almost every organization is increasingly concerned with capturing the consumer's attention through their social media content. With traditional television and radio, it's difficult to create a corpus of the content and even harder to measure the audience's reaction. With these digital platforms, we can take advantage of the digitized content and metrics such as "retweets," "shares," and "likes." This paper aims to analyze a subset of this online content by understanding how topic modeling techniques can be applied to Facebook page data. It also seeks to understand which of these topics are more popular than others.

2 Data Collection

The data for this study was collected by accessing the Facebook API (Application Programming Interface). Through making a call, or request to the Facebook API, developers are able to access information about posts on public pages including what time the post was created, what type of post it was (photo video, status update, or event), the text of the post, and how many likes/comments/shares it has received. There are some limitations to this data collection however. For instance, developers aren't able to search through all the posts on Facebook with keywords.

2.1 Data Cleaning

In social media data, it's common to find emoticons in the text. For this analysis, emoticons were deleted from all posts.

Secondly, stopwords were removed from the corpuses. This list of stopwords is based on the Ranks.nl[1] webpage and contains 174 words.

3 Methods

The primary method used in this study is MALLET[2] topic modeling. This technique starts by taking each word in the corpus and randomly assigning it to a topic. Then it goes through the following steps a

[1] Stopwords List: https://www.ranks.nl/stopwords

[2] MALLET Topic Modeling: http://mallet.cs.umass.edu/topics.php

specified number of iterations: For each document, for each word, the algorithm multiplies the probability the document contains those topics already by the probability a topic contains that word and assigns a new topic to the word with those probabilities[3]. This study uses 1000 iterations to form topic models.

Through this method, there are a few variables that will heavily influence the quality of these topics, specifically:

1. The stop words chosen.
2. The definition of a "document"
3. The number of "documents" the model is trained on
4. The number of topics the model creates
5. Optimizing hyper-parameters

Reading through the topic modeling output gives a strong sense of how these variables influence the topic quality, but we can also use a few objective measures to assess topic quality.

3.1 Co-document frequencies[4]

One method of evaluating the quality of a topic is how often the top terms co-appear in the documents. High quality topics have most of their top terms appearing together. Below is an example of topic's co-appearance matrix where the diagonal values represent how many times a certain word appears in all the documents and values in the off-diagonal represent how many documents two words appear in together. More nonzeros indicates a higher quality topic. This analysis looked at the co-document frequencies for the top 5 words.

	tax	year	plan	cut	want
tax	23	2	5	5	3
year	2	72	4	0	0
plan	5	4	34	2	0
cut	5	0	2	21	0
want	3	0	0	0	12

3.2 Topic Appearance Frequency

Another method of evaluating the quality of a topic is to see how often the topic appears in the documents. If it appears frequently and in small amounts, then the topic is very general and usually uninteresting. If it appears infrequently and in large amounts, documents, then the topic is more focused.

3.3 Stopword Selection

Even after the 174 stopwords supplied by Rank.ml were removed from the corpus, many trivial words found their way to a topic's "top words" and produced meaningless topics. Approximately 40 of these trivial words were removed from the

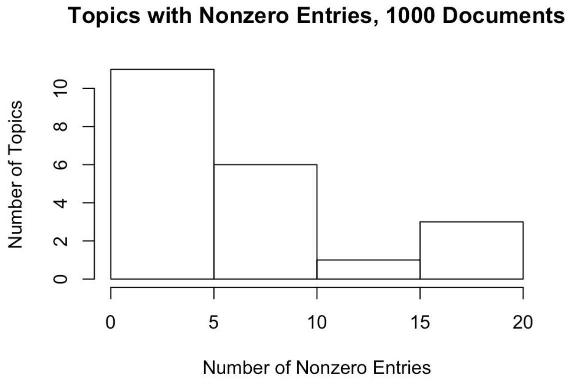
[3] Further explanation on MALLET Topic Modeling https://bigsonata.wordpress.com/2015/03/15/automatic-topic-modelling-with-latent-dirichlet-allocation/

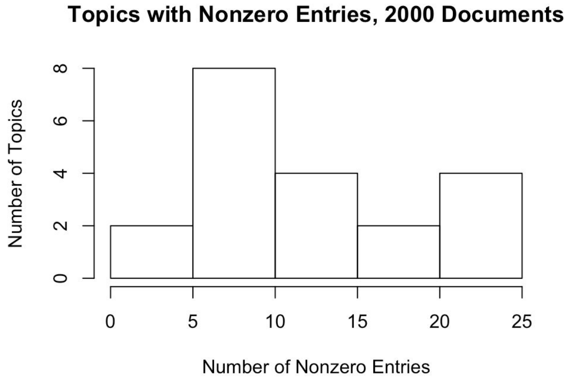
[4] Co-document frequencies https://bigsonata.wordpress.com/2015/03/15/automatic-topic-modelling-with-latent-dirichlet-allocation/

corpus, for example, “there’s,” “will,” “also,” etc.

3.4 Number of Training Documents

In all Facebook page corpuses, increasing the number of documents for the topic model to be trained on yielded dramatically better results. Of course, this also makes intuitive sense. The histograms below compare how many topics had nonzero entries when trained with 1000 documents and 2000 documents on data from Facebook’s “Business Insider” page.

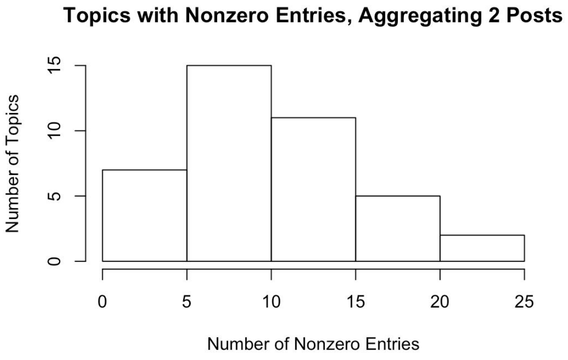


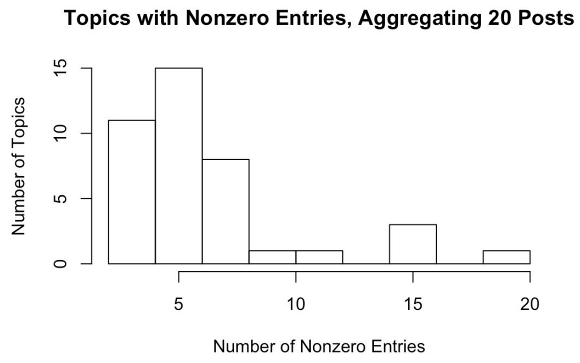


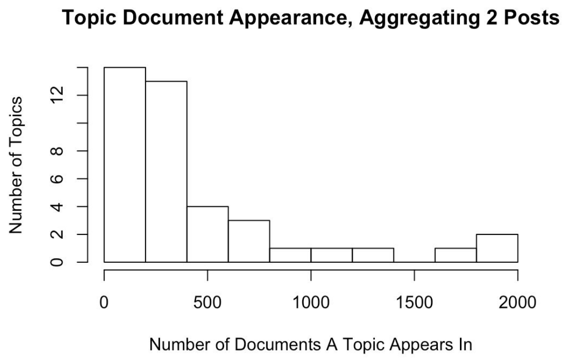
Computational power limits the number of documents retrieved to a few thousand. This study uses 2,000 documents for the analysis.

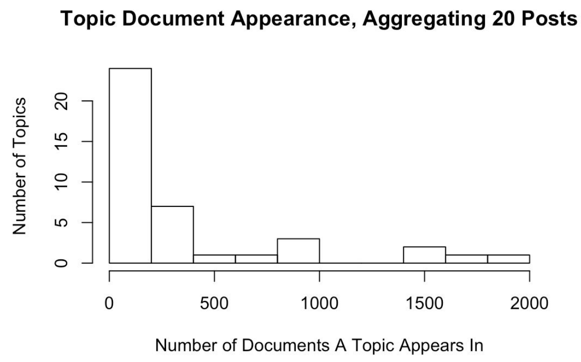
4.5 Definition of a Document

This analysis has referred to a “document” as 1 Facebook post. For news pages such as “Business Insider,” 1 post averages 50 characters in length, but for pages featuring human interviews, such as “Humans Of New York,” 1 post average 500 characters in length. Topic modelling techniques assume that each document has multiple topics. When the document is short as in the case of “Business Insider,” the model is less likely to do well. For this particular Facebook page, the study experimented with aggregating the posts to create a longer document and discovered that aggregating 2 posts produces the best results. For the longer “Humans Of New York” documents, the model was able to create quality results and did not benefit from aggregating.









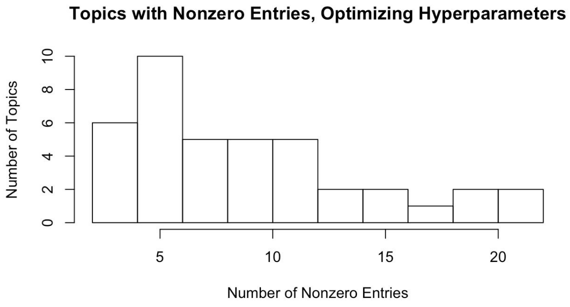
Though aggregating 20 posts were able to produce more specific topics, ultimately these topics were inferior to the topics produced from aggregating 2 posts due to the frequent zero entries in their co-document matrix.

3.6 Number of Topics

Experimentation with number of topics showed that choosing 30-50 topics created the highest quality output. Setting the topic number to 100 created topics that seemed to combine multiple themes: "parents" "success" "age" "taylor" "swift" "olympic" "won" "children" "retire" "lawsuit." Choosing too few topics like 10 seemed to miss a lot of the content.

3.7 Hyperparameter Optimization

This study also experimented with optimizing hyperparameters. Below, the hyperparameters are optimized every 20 iterations after 50 burn iterations. Overall, optimizing hyperparameters resulted in poorer topics.



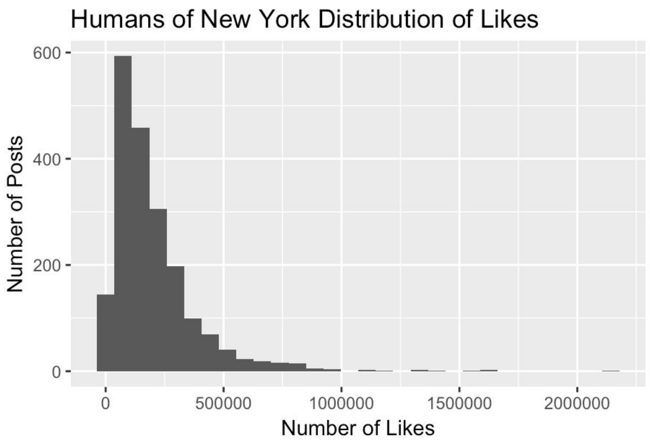
4 Assessing Topic Popularity

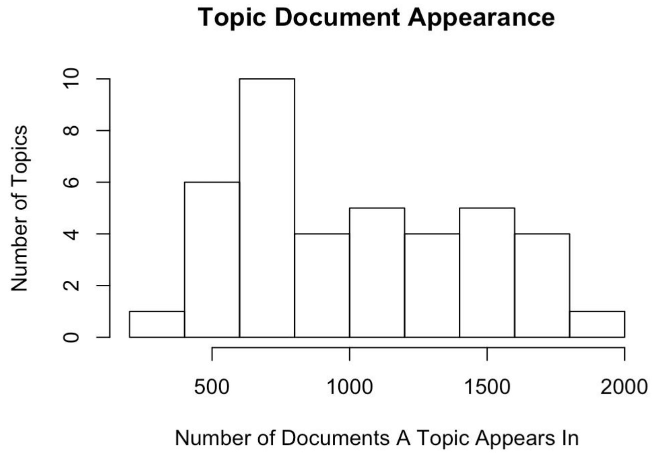
After experimenting with different stopwords, training documents, document aggregates, number of topics, and hyperparameter optimization, different topics were created for the “Humans of New York” and “Business Insider” Facebook page. Next, this study analyzes the data from these pages. Similar analyses can be reproduced with other Facebook pages.

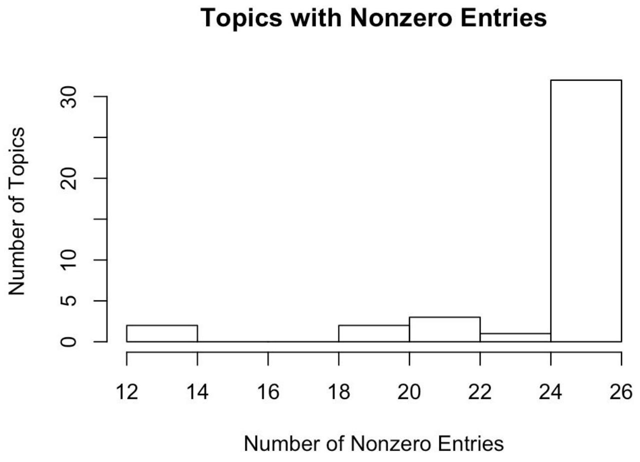
4.1 Humans of New York

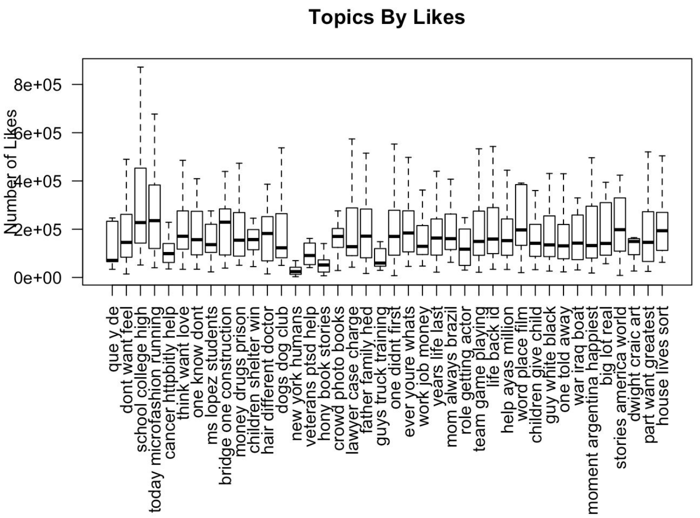
The engagement on the Humans of New York posts were heavily skewed. Most posts got between 0 and 500,000 likes. The topics created ranged from very general to very specific and topics co-appeared in documents frequently.

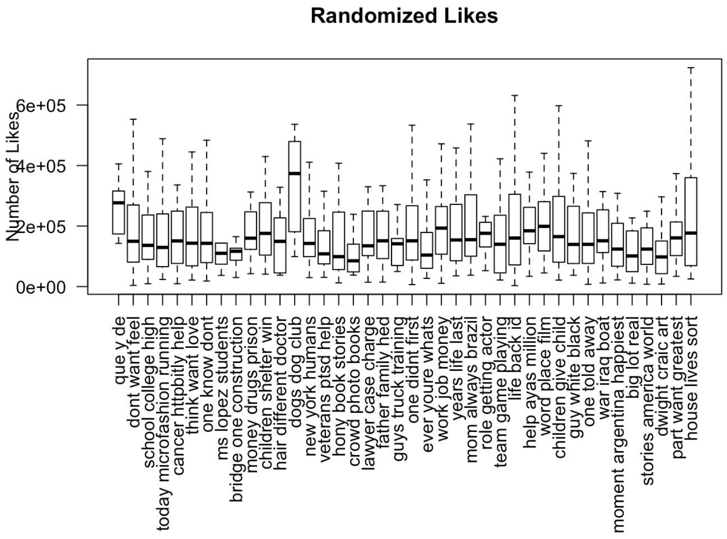
The figure on the next page compares different topics by how many likes they received. This was calculated by taking the topic that appeared most frequently in the document, and awarding that document's "likes" to that topic. On the right, the likes are randomly distributed among the topics and you can see that the boxplots have less variation, though not incredibly obvious. The two topics that received the least "likes" had the top words "new" "york" "humans," and top words "hony" "book" "stories" (hony being an acronym for Humans of New York), suggesting that self-promoting posts were not as well-received. On the other hand, the topic that performed the best had the top words "school" "college" "high." Given that most Humans of New York readers are in high school or college, this suggests that the audience enjoys reading about people in their same age demographic. "Today" "microfashion" "running" also performed well while "cancer" "httpbitly" "help" did not. See the next page for these figures.





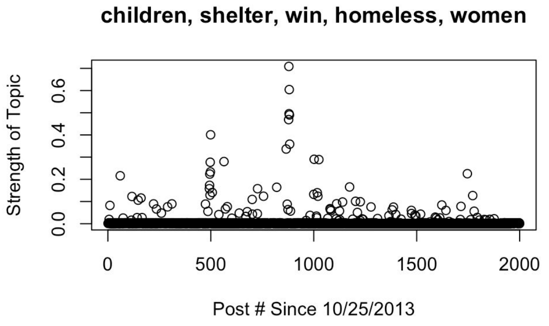


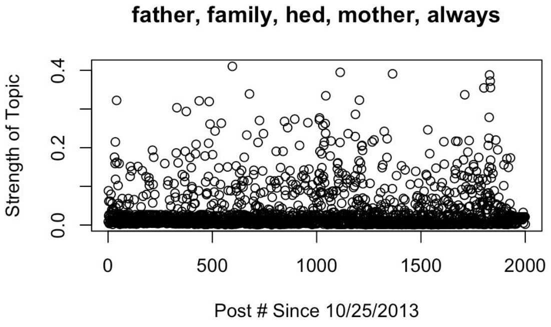


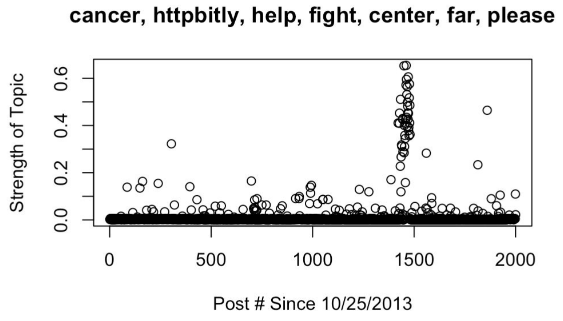


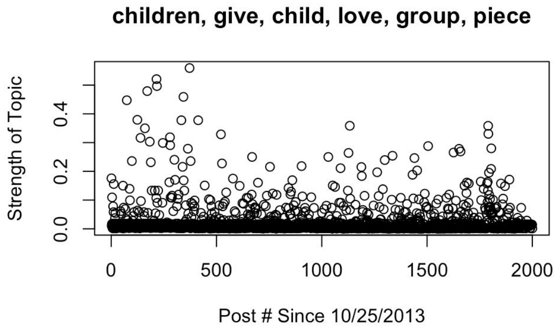
These topics were also plotted over time and showed certain patterns. More negative themes were posted about sporadically, while happier themes were posted about more continuously.

Post #500, #1000, and #1500 occur on 08/22/2014, and 07/20/2015, 06/19/2016 respectively.

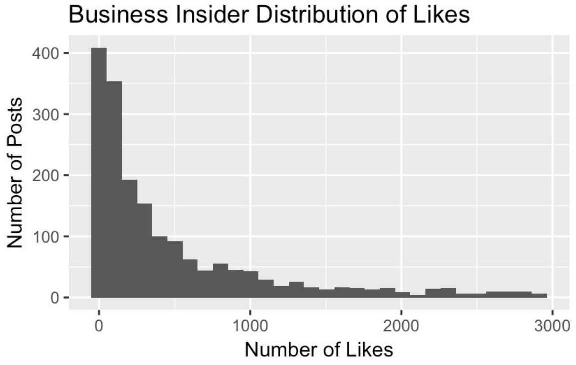








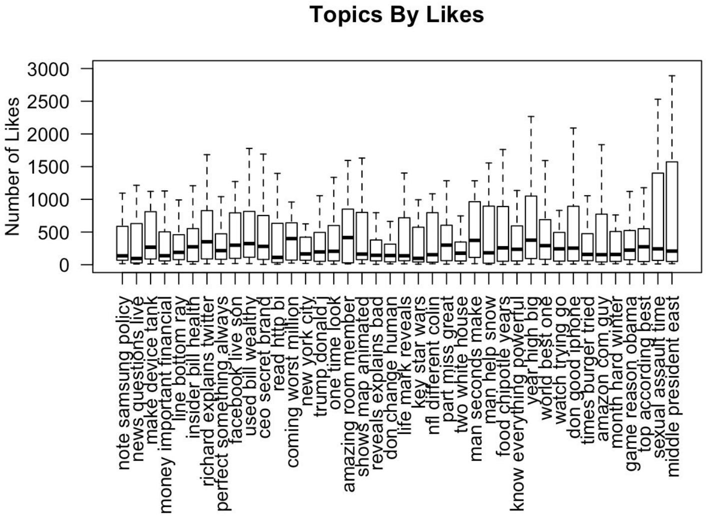
4.2 Business Insider

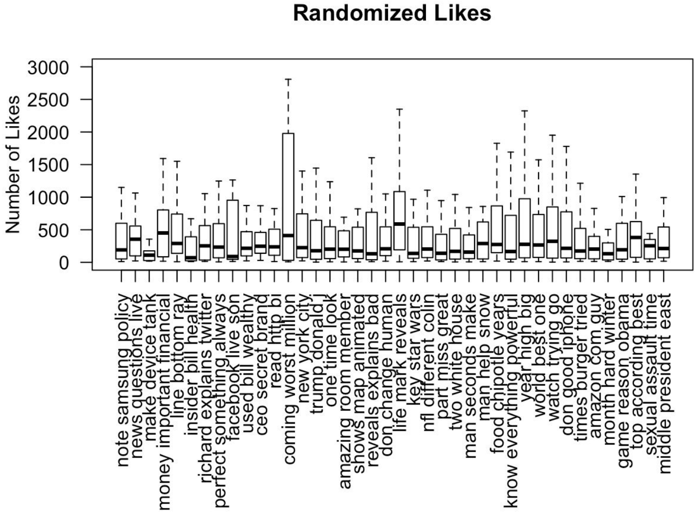


Business Insider data was chosen as a focal point due to the variety of topics they cover online; however it presented a much larger challenge due to the short document length. The topic quality was much poorer, as can be seen from the Figures in Section 2. Though initially, the topics seemed to produce interesting results, after randomizing the "likes," it becomes obvious that the variation in "likes" could have occurred by chance, therefore the analysis wasn't able to conclude anything meaningful from the topic popularity comparison.

5 Conclusion and Discussion

In conclusion, we can use MALLET topic modelling to break down Facebook posts into topics with varying degrees of success. Increasing the training size and using longer document lengths greatly improves the quality of the topics, while aggregating multiple documents into one isn't as successful. We can further suggest that self-promoting posts and more negative posts don't perform as well as some relatable and trendy posts. For further investigation, instead of looking at the "likes" count we could also analyze the shares, comments, and different reactions on Facebook. With these further studies, we can better understand how audiences are consuming their content online to help shape the digital communication strategies.


Topics By Likes
Number of Likes
note samsung policy
news questions live
make device tank
money important financial
line bottom ray
insider bill health
richard explains twitter
perfect something always
facebook live son
used bill wealthy
ceo secret brand
read http bi
coming worst million
new york city
trump donald j
one time look
amazing room member
shows map animated
reveals explains bad
don change human
life mark reveals
key star wars
nfl different colin
part miss great
two white house
man seconds make
man help snow
food chipotle years
know everything powerful
year high big
world best one
watch trying go
don good iphone
times burger tried
amazon com guy
month hard winter
game reason obama
top according best
sexual assault time
middle president east


Randomized Likes
Number of Likes
note samsung policy
news questions live
make device tank
money important financial
line bottom ray
insider bill health
richard explains twitter
perfect something always
facebook live son
used bill wealthy
ceo secret brand
read http bi
coming worst million
new york city
trump donald j
one time look
amazing room member
shows map animated
reveals explains bad
don change human
life mark reveals
key star wars
nfl different colin
part miss great
two white house
man seconds make
man help snow
food chipotle years
know everything powerful
year high big
world best one
watch trying go
don good iphone
times burger tried
amazon com guy
month hard winter
game reason obama
top according best
sexual assault time
middle president east