Facebook Sentiment Analysis (Data Mining)

EN.550.436 Data Mining

Jeong eun (Hailey) Lee

Alex Ahn

Jae Goan Park

Overview

Sentiment Analysis is more than figuring out how people feel about in the social media. With sophisticated analysis on how people react to certain topics, sentiment analysis can predict the following: campaign success, marketing strategy, product messaging, customer service, and stock market price.

We decided to take advantages of recent extension of reactions made by Facebook and do sentiment analysis on how people react differently for different posts.

Our dataset consist of all posts from "Opposing View" from Facebook Public page from August 2016 to April 2017.

Objective

- **Extract data from Facebook Public Page**
- **Label topics using LDA (Latent Dirichlet Allocation)**
- **Visualize Facebook statuses using PCA**
- **Unsupervised Learning: assign most relevant topic to each status**
- **Supervised Learning: using labels we learned that we can classify topic**

Method

1. **Select a public Facebook page for user sentiment analysis**
2. **Collect data including status and user responses**
3. **Perform LDA (Topic Modeling algorith, Latent Dirichile Allocation on status corpus and generate topic vectors**
4. **Assign most relevant topic label to each status (based on highest probability**
5. **Perform PCA to visualize topic distribution**
6. **Visualize correlation matrix of the topic distribution**
7. **Visualize user responses to each topic**
8. **Cluster status corpus using topic-features (KMeans, GMM)**
9. **Plot likelihood/inertia for each K-number of clusters for each method**

Topic Distribution in LDA model

1. **Topic 1 (Family, Urgent):**

   walked didn time way man make decided called immediately getting wife realized baby away mom did hours rushed hospital boyfriend

2. **Topic 2 (Offence):**

   woman man young words knew began didn judge stop suddenly got went took felt inside later received way home offended

3. **Topic 3 (School Crime):**

   old year girl boy school man mom little mother home just daughter happened police father got years son like killed

4. **Topic 4 (Police/Crime/Satire):**

   police car officer woman saw wasn life pulled just officers left door arrived won couple looked hard cop driver soon

5. **Topic 5 (Donald Trump):**

   trump president donald obama just america think agree like says support new good news disagree right bad americans michelle office

6. **Topic 6 (Teenage Abortion):**

   baby thought woman decide teen girl told just doctors man wrong took look decided right went gave did doctor teacher

7. **Topic 7 (Marriage):**

   husband day decided years dog saw national wife hands anthem worked finally kaepernick secret like colin vote come having stand

8. **Topic 8 (Crime Witness):**

   realized people noticed look took going saw closer picture house doing photo couldn quickly thing lot viral white suddenly walmart

9. **Topic 9 (Attack/Terrorism):**

   didn message world children officials people think elderly trying wasn say things just started long isis home expecting does brutal

10. **Topic 10 (Hillary Clinton):**

    clinton hillary just worse big doing son death entire mom simple question time sex worst going got woman did ll

Conclusion

Able to observe/tracked distributions of different topics under "Opposing View" in the facebook group page.

We chose 10 topics to be able to interpret the results more clearly.

From August 2016 to April 2017, as we can see from the graph, Donald Trump took the majority of attention, especially because it is politically highly debated topic.

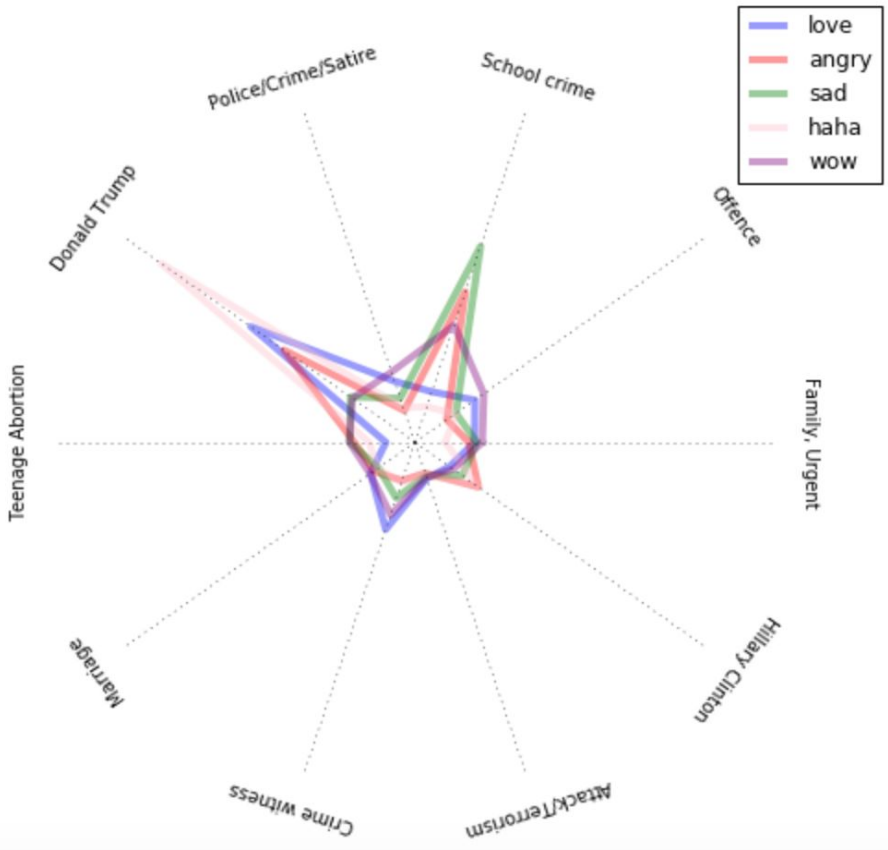
From clustering, (both K-means and GMM), we observed that likelihood approximately converges after 10 cluster mean. This means that topics are usually combined with one another for each status.

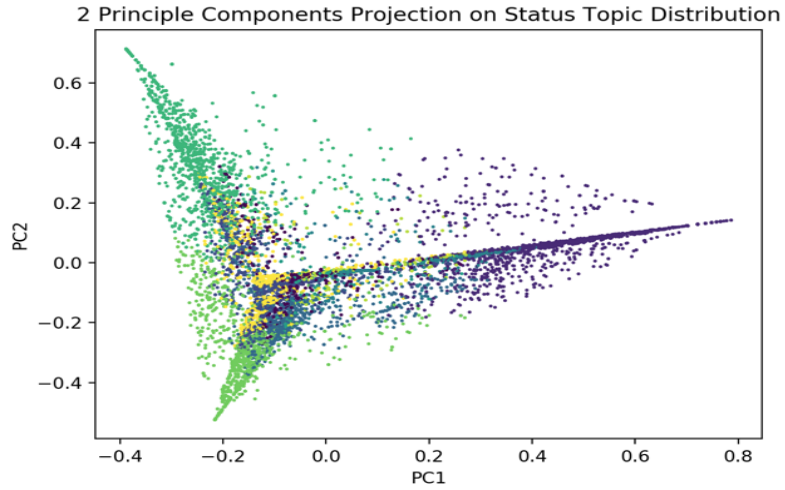
According to the correlation matrix, we noticed that Teenage abortions, Family issues, and Offense are highly correlated with one another. Police Crime Satire is likely correlated with crime witness, while Donald Trump is not correlated with anything in topic distribution of each status.

Further implications: apply to the other page, and we can possibly predict how people react sensitively to certain topics.

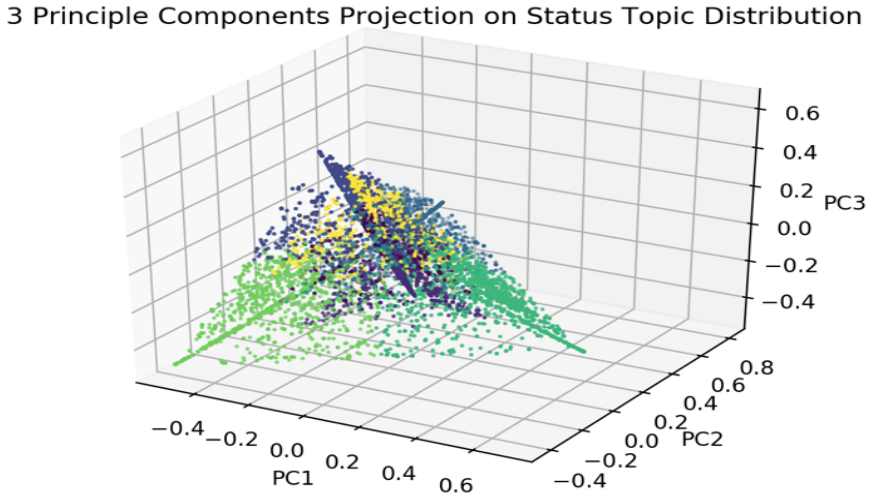
Figures

After extracting 10 set topics using LDA, we decided to plot it various ways to verify our results. First, we simply plotted the radar charts with 10 topics as circle and have 5 different emotion variables. As you can observe here, we had the most reactions, whichever, for Donald Trump. Which makes sense, since our data was collected from ...(recent data). Controversial topic. We observed that more of negative responses, sad, or angry appeared most for School crime and crime witness. After verifying this result by visualizing each emotional reaction, we went forward to compute Principal Component Analysis.





Explained variance with 2 eigen vectors: 0.390087%

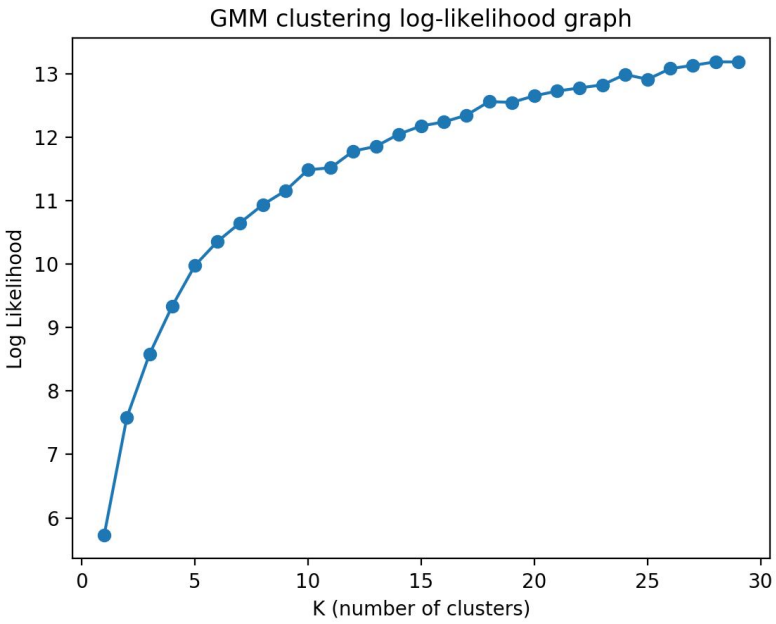


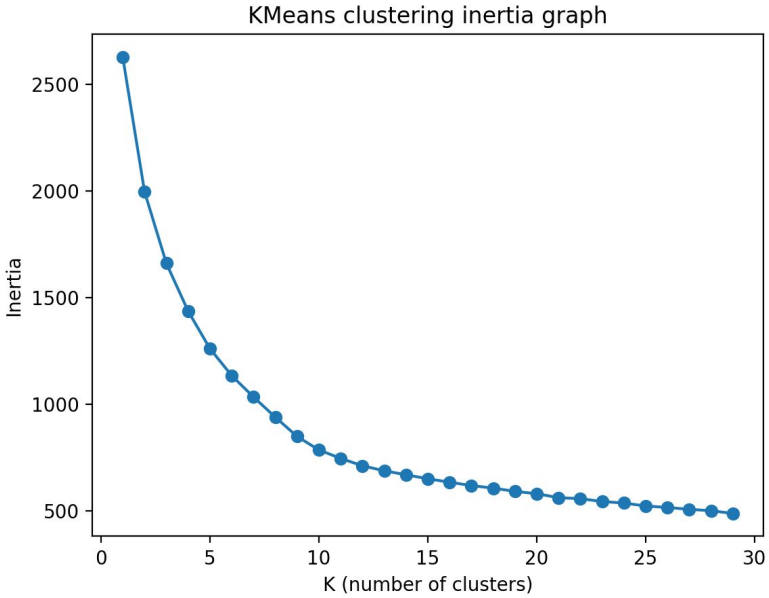
Explained variance with 3 eigen vectors: 0.531836%

So each document vector equals one topi.

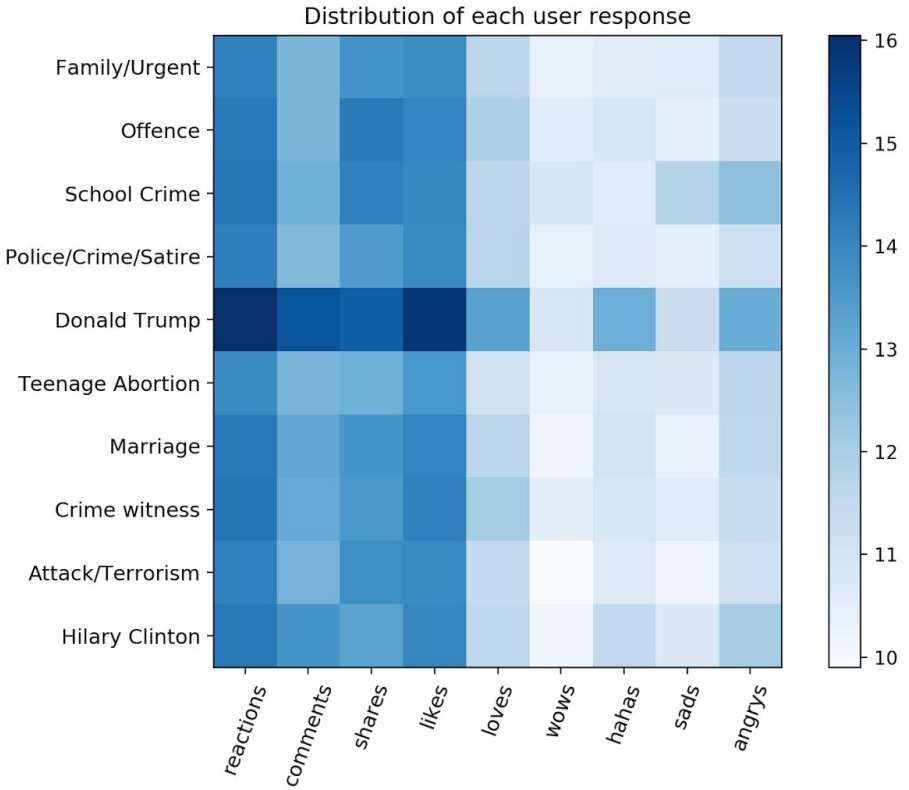
2dimensional, 3 dimensional , eigenvector = set of document topic distribution, lower dimensinoal, hard to distinguish 10 features, Topic distribution sums up of 1.

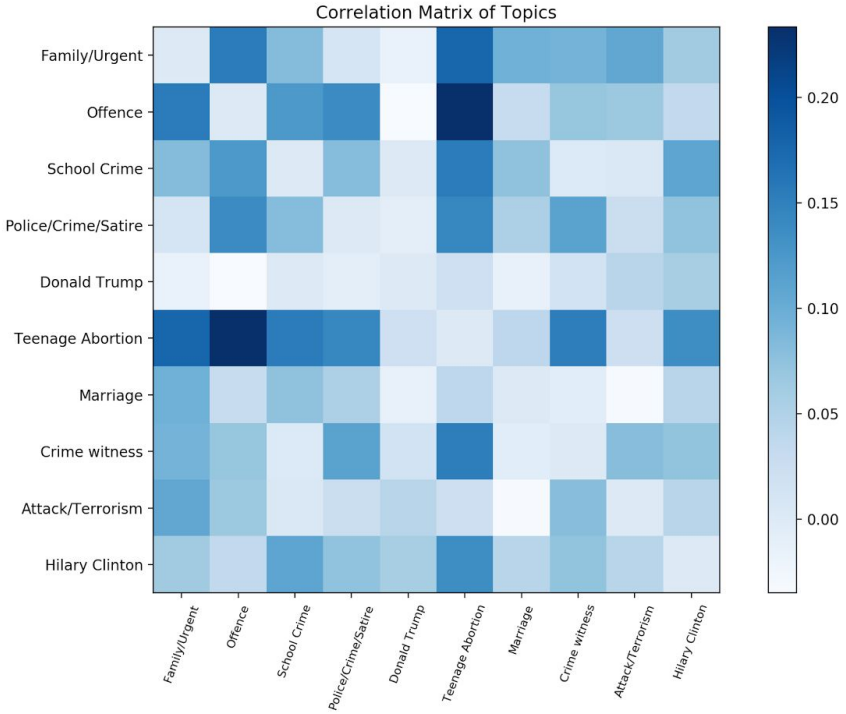
Topic distribution by eigenvector, wanted to visualize the data points, document vector== topic distrubtions





we could confirm 10


Distribution of each user response
Family/Urgent
Offence
School Crime
Police/Crime/Satire
Donald Trump
Teenage Abortion
Marriage
Crime witness
Attack/Terrorism
Hilary Clinton
reactions
comments
shares
likes
loves
wows
hahas
sads
angrys
16
15
14
13
12
11
10


Correlation Matrix of Topics
Family/Urgent
Offence
School Crime
Police/Crime/Satire
Donald Trump
Teenage Abortion
Marriage
Crime witness
Attack/Terrorism
Hilary Clinton
0.20
0.15
0.10
0.05
0.00

References

https://github.com/minimaxir/facebook-page-post-scraper

https://medium.com/@baditaflorin/understanding-facebook-reactions-using-sentiment-analysis-f17b6e561ff3