Kexin Guan, Peter Min

Dr. Lauren Klein

QTM 340

12/15/2019

QTM 340 Final Project

I. Introduction

The aim of this project is to examine the newspaper articles published by *The Emory Wheel*, an independent, student-run newspaper based at Emory University in Atlanta, Georgia. Primarily, we would like to examine the range of topics covered by the newspaper, as well as the existence of possible cyclic patterns behind these topics. The rationale for the project is that while the range of topics covered by Emory Wheel tends to be wide when it is measured in terms of the traditional 'news', we believe that at the same time, the discussion of such topics appear to be unbalanced for the reason that more focus is put into specific subjects--for example, U.S. political events and Student Government Association--with possible patterns in the language and timing that the news is reported. Therefore, by examining the topics covered by this newspaper and the patterns behind them, we hope to provide some insights on how the independent media is reflecting the events around us, as well as coming up with possible improvements to the news coverage so that students can be more interested in learning about what is happening in the world and more importantly, guiding the Emory Wheel to become a more popular news source.

II. Related Research

CS224N Final Project: Sentiment analysis of news articles for financial signal prediction

The first project that we chose involves a final project of a class at Stanford University carried out by three students. These students focused on the analysis of publicly-available news reports with the use of computers to provide advice to traders for stock trading. Our project is related to their with regard to the usage of sentiment analysis on news reports. However, it is also different from our project in that it extracts sentiment from a natural language for the purpose of providing it as an input to that larger system, while we will mainly focus on the analysis of the result. More importantly, we use simple built-in functions while they both built classification themselves and used an automatic classifier.

Using Text Analysis in R to Analyze News – Part 1 - Loretta C. Duckworth Scholars Studio

The second project we examined applied the R programming language to conduct simple text analysis. It relates to our project in a way that both of our projects are intended to analyze news reports. Nevertheless, it is also different in a few ways. First, the author of this project employs R as the main tool for analysis and he uses a very large database to obtain his target corpus. Secondly, it only looks at news reports related to the LGBTQ community while we intend to examine the entire archive of *The Emory Wheel*. Lastly, the most important distinction would be that the author intends to examine the coverage for a minority group reported in the news, while we intend to examine the focus and the sentiment of the news publishers themselves.

U.S. journalism really has become more subjective and personal — at least some of it

The last report presents a quantitative assessment of how the presentation of news has changed over the past 30 years and how it varies across platforms. It is related to our project in

terms of the comparison of media contents across time, as well as the study for subjectivity. However, while this study focuses on a wide range of media across different platforms, we only focus on a single newspaper in a relatively shorter time.

Based on these three projects, it can be seen that our combination of sentiment analysis and time factors complements the second project by adding a time factor as well as providing a finished prospect since the author never actually released the following posts for his analysis. Our choice of a student-run school newspaper also stands out as a distinctive factor that could possibly fill in a blank spot in the research area where data sources are dominated by voluminous literature or scientific paper archives. We believe, if possible, our project is able to contribute to a larger picture where the change of sentiments over time among United States college students would be examined.

III. Corpus

The corpus used for this final project is a collection of all the newspaper articles published by *The Emory Wheel* ranging from October 2nd, 2014 to October 1st, 2019. The articles were first manually downloaded in PDF versions from the Newsbank website using Emory credentials. They were then converted into text file formats using online conversion tools.

From here, the articles, which are now text files, are read into a Jupyter notebook and were first filtered using regular expressions to get rid of any unnecessary words or sentences that might pollute the corpus such as ‘Copyright © 2017’, ‘WordsOpenURL Link’. Then, complex regular expressions and control flows were applied to partition the components of each article as

accurately as possible. Finally, the components were recorded into a pandas dataframe which has four columns specifying the title, author, date and detailed content for each article.

IV. Process and Methods

The methods selected to process the dataset include an examination of the TF-IDF scores, word counting, sentiment analysis, and finally topic modeling and the associated visualization using pyLDAvis. The first method that we tried was to examine the TF-IDF scores and word counting. However, it turned out that regardless of the year that we examined, words such as 'emory', 'university', 'students' always predominate. Similarly, neither did word counting return us any interesting results since the results for these two analyses tend to be highly correlated.

The second method that we applied was topic modeling. We applied topic modeling first to the entire corpus to gain a basic understanding of the topics covered by the newspaper. Then, we further examined the topics covered at the beginning and the end of the semesters by defining the beginning as September and February, the end as May and December.

Lastly, we used sentiment analysis to calculate an aggregated value for each month for a total of 61 values. Meanwhile, the date column was transformed into a Python datetime format which was sorted as ordinal variables. These two factors of interest were eventually combined together to examine the change and trend of the newspaper's sentiments over time.

V. Results and Discussion

For the result of topic modeling for the entire corpus, we got the 15 topics as our result:

```
T0: sga, said, elections, palmer, votes, voting, election, cohen, percent, president, board, ballot, runoff, ma, candidates,
vote, ox, students, wheel, email,
T1: emory, mental, wheel, showcase, health, arts, performances, event, editorial, post, disorders, students, dark, therapist,
poetry, stories, included, reporting, board, mark,
T2: emory, students, year, kaldi, plan, university, percent, according, campus, master, food, said, admissions, college, health,
atlanta, new, admitted, wheel, process,
T3: israel, jewish, israeli, menu, wall, meal, chicken, esjp, options, palestinian, palestine, palestinians, ordered, conflict,
anti, market, pro, hot, served, tea,
T4: spc, dooley, sodexo, workers, week, according, bon, ebola, agency, employees, church, cole, concert, virus, homecoming,
disease, patients, migos, contract, sex,
T5: trump, according, president, government, state, senate, party, march, states, investigation, people, united, vote,
political, public, rights, election, georgia, country, campaign,
T6: involved, new, league, country, football, super, week, nfl, los, california, sports, pitch, angeles, york, scott, players,
correspondent, quarterback, win, come,
T7: song, music, gray, album, songs, audience, night, crowd, pop, pizzeria, set, stage, culture, artists, sound, concert, tour,
atlanta, crush, band,
T8: team, place, emory, women, men, second, epd, said, yard, time, meet, event, freestyle, individual, eagles, senior, finish,
title, championship, freshman,
T9: said, kavanaugh, emory, women, singles, win, doubles, match, rally, mora, allegations, court, matches, eagles, sexual,
victory, season, judge, williams, college,
T10: emory, said, students, college, university, according, campus, people, student, community, school, life, year, new,
members, president, atlanta, work, wheel, program,
T11: student, committee, sga, council, president, said, long, students, saf, prasad, cc, gsga, graduate, vice, government,
university, honor, course, funding, wheel,
T12: film, like, time, world, think, know, best, way, people, story, feel, new, life, love, day, ll, years, good, ve, movie,
T13: team, game, emory, eagles, season, said, university, second, points, win, senior, junior, play, year, games, sophomore,
time, teams, lead, goal,
T14: film, vegan, pizza, episode, chef, cheese, times, food, death, episodes, robots, characters, love, bad, peele, time, like,
meat, dough, animation,
```

As we can observe from the result above, the major topics of *The Emory Wheel* include American politics, Student Government Association, campus life, national sports, Emory sports team, the conflict between Palestinian and Israeli ideologies, Atlanta lifestyles and various dining options.

While the overall results reflect the typical topics that *Emory Wheel* covers in its publishings, we were also interested in how the topics would be at the start and end of each semester because at the beginning of each semester, one could expect the topics of the newspaper to be covering the elections of new Student Government Association and campus life as new faces are coming to Emory. Meanwhile, at the end of each semester, one could expect an increase in the news coverage of finals or graduations. Here are the topics covered at the start and end of each semester:

At the start of the semester:

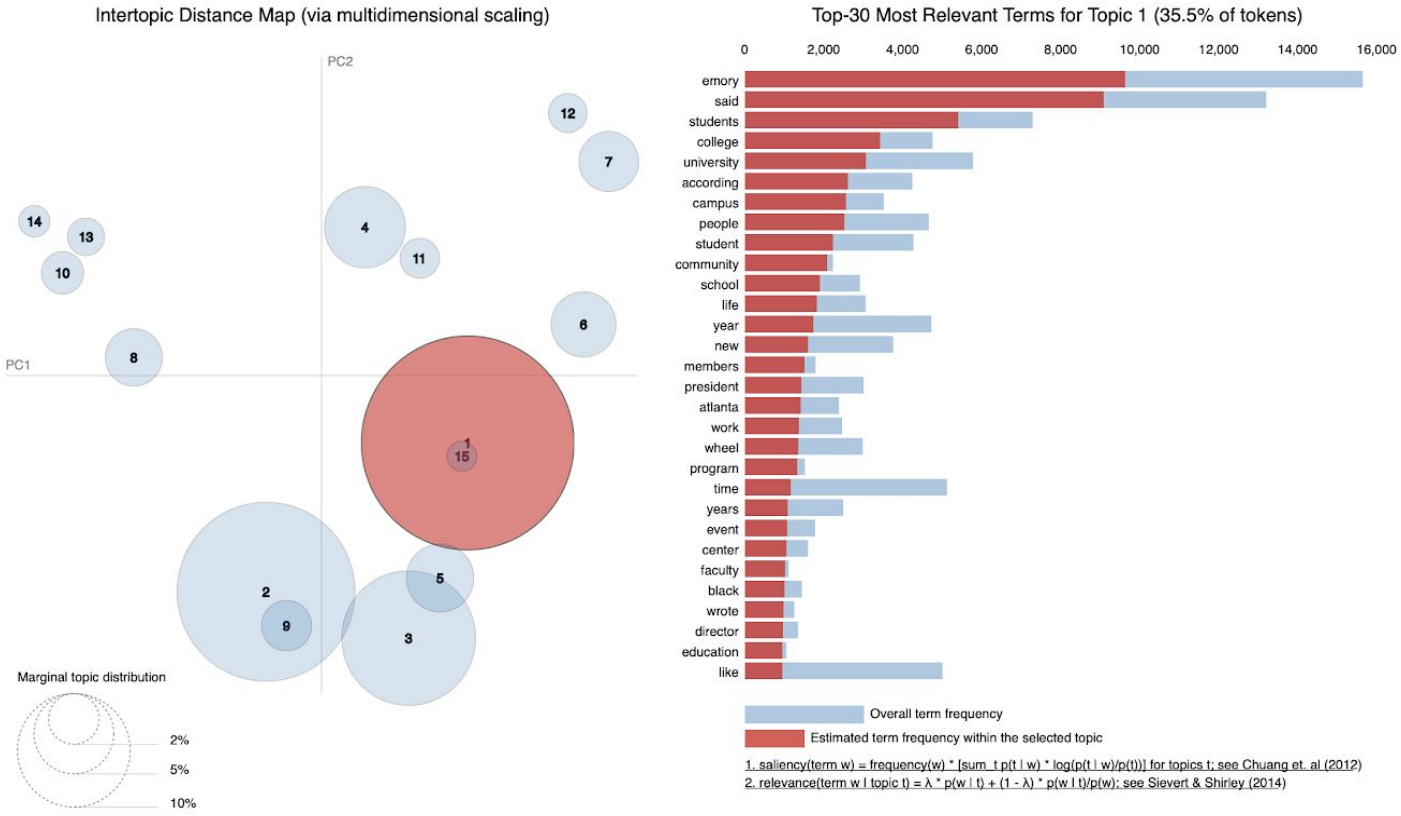
```
T1: 3.01% of document.
Top words in topic: spc, dooley, sodexo, workers, week, according, bon, ebola, agency, employees, church,
cole, concert, virus, homecoming, disease, patients, migos, contract, sex,
T2: 7.12% of document.
Top words in topic: trump, according, president, government, state, senate, party, march, states,
investigation, people, united, vote, political, public, rights, election, georgia, country, campaign,
T3: 1.72% of document.
Top words in topic: involved, new, league, country, football, super, week, nfl, los, california, sports,
pitch, angeles, york, scott, players, correspondent, quarterback, win, come,
T4: 3.16% of document.
Top words in topic: song, music, gray, album, songs, audience, night, crowd, pop, pizzeria, set, stage,
culture, artists, sound, concert, tour, atlanta, crush, band,
T5: 27.89% of document.
Top words in topic: emory, said, students, college, university, according, campus, people, student,
community, school, life, year, new, members, president, atlanta, work, wheel, program,
T6: 5.69% of document.
Top words in topic: student, committee, sga, council, president, said, long, students, saf, prasad, cc,
gsga, graduate, vice, government, university, honor, course, funding, wheel,
T7: 36.73% of document.
Top words in topic: film, like, time, world, think, know, best, way, people, story, feel, new, life,
love, day, ll, years, good, ve, movie,
T8: 8.86% of document.
Top words in topic: team, game, emory, eagles, season, said, university, second, points, win, senior,
junior, play, year, games, sophomore, time, teams, lead, goal,
T9: 3.61% of document.
Top words in topic: film, vegan, pizza, episode, chef, cheese, times, food, death, episodes, robots,
characters, love, bad, peele, time, like, meat, dough, animation,
```

At the end of the semesters:

```
T1: 23.72% of document.
Top words in topic: emory, mental, wheel, showcase, health, arts, performances, event, editorial, post,
disorders, students, dark, therapist, poetry, stories, included, reporting, board, mark,
T2: 34.50% of document.
Top words in topic: emory, students, year, kaldi, plan, university, percent, according, campus, master,
food, said, admissions, college, health, atlanta, new, admitted, wheel, process,
T3: 14.05% of document.
Top words in topic: israel, jewish, israeli, menu, wall, meal, chicken, esjp, options, palestinian,
palestine, palestinians, ordered, conflict, anti, market, pro, hot, served, tea,
T4: 6.16% of document.
Top words in topic: spc, dooley, sodexo, workers, week, according, bon, ebola, agency, employees, church,
cole, concert, virus, homecoming, disease, patients, migos, contract, sex,
T5: 1.43% of document.
Top words in topic: trump, according, president, government, state, senate, party, march, states,
investigation, people, united, vote, political, public, rights, election, georgia, country, campaign,
T7: 5.16% of document.
Top words in topic: song, music, gray, album, songs, audience, night, crowd, pop, pizzeria, set, stage,
culture, artists, sound, concert, tour, atlanta, crush, band,
T8: 6.56% of document.
Top words in topic: team, place, emory, women, men, second, epd, said, yard, time, meet, event,
freestyle, individual, eagles, senior, finish, title, championship, freshman,
T9: 8.06% of document.
Top words in topic: said, kavanaugh, emory, women, singles, win, doubles, match, rally, mora,
allegations, court, matches, eagles, sexual, victory, season, judge, williams, college,
```

As we can see from the results, at the beginning of all the semesters, the newspaper tends to cover topics such as extracurricular activities, which can be seen in the 36.7% value placed on T7 as well as life as a new Emory student, represented by a 27.89% value placed on T5. Meanwhile, for the end of the semesters, Topic 1 rises up to a value of 23.72% and it reflects the stress that students face near the end of the semesters when they are dealing with exams. T2, existing within 34.05% of all the end-of-the-semester documents, also correctly reflects a trend that is symbolic, as the discussions on admission and everything revolving around Emory sprouts when semesters end. However, the 14.05% value placed on T3 should not be represented since the ideological conflict between Israeli and Palestinians only occurred for one semester before university officials put a halt to it.

Lastly, before we move on to sentiment analysis, here is one of the visualizations of the topics modeling result using pyLDAvis, here the relevance value λ is set to 1.



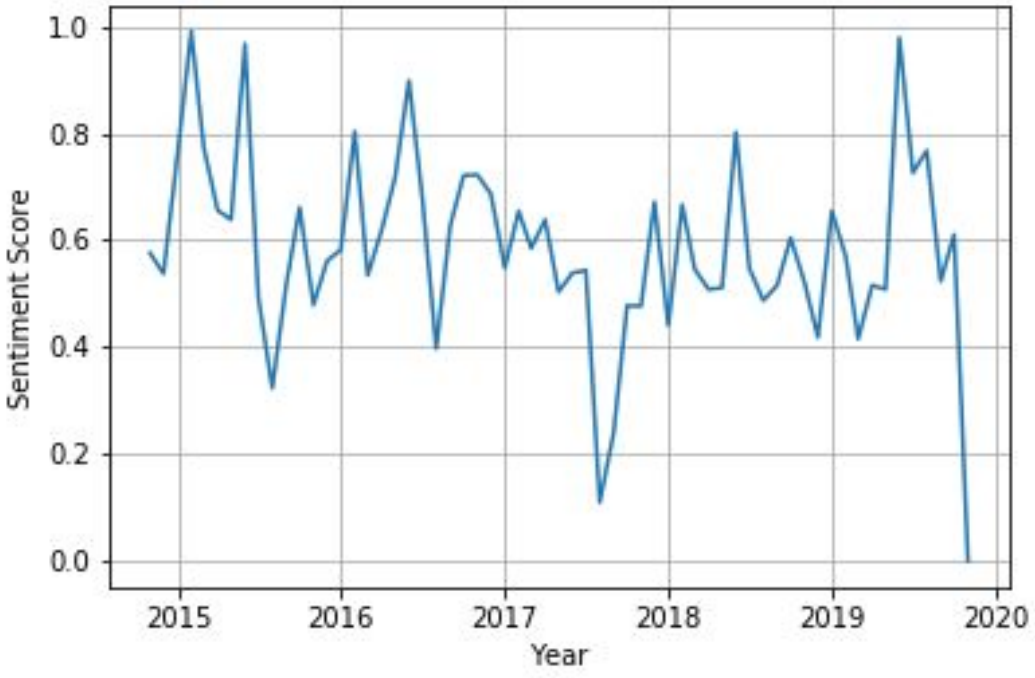
The last method that we employed was sentiment analysis for the corpus following the time order. Here we obtained the results for the news reports with the highest and lowest sentiment scores among the entire corpus.

Title	Author	Date	Content	Compound
Make War On Rapists, Not Drugs	Safiyah Bharwani	2015-09-20	Trigger Warning: Sexual AssaultFor most victim...	-0.9998
The Barkley Forum Debates	Emory Wheel	2015-09-02	Barkley Forum members Katie Duval (sophomore) ...	-0.9997
Barkley Forum Debates: The Confederate Flag	Emory Wheel	2015-06-29	Given the recent controversy surrounding the f...	-0.9996
Jews Must Back Gun Control	Jacob Busch	2018-11-07	More than 10,000 hate crimes involving a firea...	-0.9995
Aquinas Center Hosts Death Penalty Debate	Emily Sullivan	2016-02-03	Emory's Aquinas Center of Theology pitted two ...	-0.9995
...	...	...	...	...
Racial Justice Retreat Addresses Demands	Karishma Mehrotra	2016-03-01	Administrators, faculty, staff and student act...	0.9998
Pitching Your Voice at Emory: Behind the Scene...	Olivia Shuler	2015-10-16	You thought Ivy League acceptance rates were s...	0.9998
Best Buddies Builds Bonds	Lindsay Wilson	2015-03-26	Once a month, Best Buddies, hosts an event whe...	0.9998
'Race' Cast, Crew Talk Storytelling, Authenticity	Brandon Wagner	2016-03-04	Courtesy of Focus FeaturesRace tells the story...	0.9999
You're Living in a Golden Age of TV Comedy	Brandon Wagner	2015-09-01	"My sense is that 2015 or 2016 will represent ...	0.9999

As we can observe from the table above, the news with the highest sentiment score includes topics such as social and racial justice, and along with them, a report about comedy sits

on the top of the list. One thing to be noticed here is that the top two articles with the highest sentiment score are composed by the same author, indicating that personal writing style of the authors may lead to a relatively strong bias in how the news is presented to the readers. As for the five posts with most negative results, the topics are centered around controversial social issues such as gun control, drug, and the death penalty. This result does not fall beyond our expectations.

At last, a graph is made to give a better presentation of the change in sentiment analysis score through time. The average sentiment score for each month is calculated, and it is plotted on the graph below.



From this figure, we can observe a slight downward trend of decrease in sentiment scores and July 2017 has the lowest sentiment score. Considering the aforementioned fact about topics

exhibiting abnormal sentiment analysis, we think that this might be associated with the effect that the inauguration and the ensuing news regarding President Donald Trump have on the editors' commentary.

In general, our results complicated our initial analyses compared to existing literature. Our combination of sentiment analysis and time series factor create a place for improvement but also leaves us with interesting results. Moreover, it left us space for future researches to be conducted since we have not yet found out a legitimate reason for the drop in sentiment score in July 2017 and more importantly, the big trend of gradually decreasing sentiment scores.

VI. Conclusion and Next Steps

Based on our analyses, it turns out that *Emory Wheel* has an extensive range of topics and the duality of diversity and singularity aforementioned in the introduction is confirmed. From here, we can advise staff and fellow students working in the newspaper that while they should maintain this extensive range of currently existing topics, they can also try to cover something different, for example, news about the student composition and how international students are composed of a crucial part in the Emory community, etc. However, our project is also limited in a few ways. First, due to the nature of this analysis, the time span of the corpus is limited to the last five years while in fact the earliest *Emory Wheel* articles archived on the Newsbank website can be traced back to 2002. By excluding all the articles from 2002 to September 2014, we chose efficiency and relevance in accordance with our identities as Emory's class of 2020 students but they also come at the cost of the significance of our examined topics, especially the valuable trend of the topics over a large time span, if there is any.

Secondly, we think that the corpus used can be further refined, not into a shorter time span but rather down to a specific topic or area of interest. An example would be the crime reports published on *Emory Wheel* which come from the Emory Police Department. This could serve as a decent corpus since crime reports are composed of great details including everything imaginable about incidents that actually happened. More importantly, every one of them is written by a professional organization within the Emory community, which might be less biased compared to other articles such as political commentaries written by regular students. By examining the topics of the crime reports, we can have a solid grasp of what parts of the Emory community are most vulnerable to crime incidents and what kind of crimes occur with the most frequencies, a point from which targeted surveillance and patrols can be provided.