BIA 660: Web Mining - Project Report
Sentiment Analysis of Mainstream Media sources using twitter

Siddesh Gannu *Shruti Shete* *Ashwin Pandey*

1. Introduction

Since the invention of social media, people have had the ability to post their opinions on a wide range of topics ranging from food to politics. At the forefront of this social media explosion is twitter which currently has the largest database of opinions. Our project aims to provide clarity to mainstream media sources about where they stand in relation to their audience with respect to the latest news topics. Does their sentiment about a topic match that of the average twitter user? Or are their opinions alienating consumers of their media.

Our project ranks media sources by the number of times people agreed the most with their sentiment about a topic. For example, people might have agreed the most with CNN on 7 topics while they agreed with the New York Times only 4 times. So we can say that CNN ranks above the New York Times. This adds a new dimension to recommendation systems in that news channels can be recommended based on their popularity in the twitter space among twitter users. Goal is to re-engineer the way twitter recommends news outlets to individual twitter users based on their previous tweets.

For example, below you will see that USA Today had the highest rank because people agreed with them the most on 3 topics. Whereas people agreed with other sources on only 1 or 2 topics.

2. Preliminary Literature Review

Social media networks have become a vital tool for sharing information and for influencing opinions and decision-making. Furthermore, the impact of social media on various sectors is growing. Social media is generating a large volume of sentiment rich data in the form of tweets, status updates, blog posts, comments, reviews, etc. The amount of content generated by end users is very vast for a normal user to analyze. So there is a need to implement the techniques to analyze the sentiments.

Sentiment analysis can be defined as a process that automates mining of attitudes, opinions, views and emotions from text, speech, tweets and database sources. Various papers presented the research studies on using sentiment analysis to predict the opinion by traditional media sources present on twitter in topics like politics, sports, food and many more. Some of the early and recent results on sentiment analysis of Twitter data are by Go et al. (2009), (Bermingham and Smeaton, 2010) and Pak and Paroubek (2010). Go et al. (2009) use distant learning to acquire sentiment data. They use tweets ending in positive emoticons like ":)" ":-)" as positive and negative emoticons like ":(" ":-(" as negative. In terms of feature space, they try a Unigram, Bigram model in conjunction with parts-of-speech (POS) features. They note that the unigram model outperforms all other models. Specifically, bigrams and POS features do not help. Another significant effort for sentiment classification on Twitter data is by Barbosa and Feng (2010). They use polarity predictions from three websites as noisy labels to train a model and use 1000 manually labeled tweets for tuning and another 1000 manually labeled tweets for testing. They however do not mention how they collect their test data. The research paper by Frank et al. (2019) illustrates that tweets are more prevalent on non-print media(TV) than print media(news). The sentiments for topics on security, politics and economics were generally negative, while sentiments on sports were positive. Along with this research, a paper by Yogev et al. (2021) analyses the opinion inversion in tweets by using sentiment analysis. Team developed machine-learning models to predict whether a Tweet will undergo Opinion Inversion.

These studies account for large-scale data collection from an online platform (twitter). But they do not emphasize on studying and analyzing tweets by Mainstream media and average twitter users. Hence, our focus is to study and analyze the tweets by these accounts and compare how close tweets and opinions are about that particular topic.

3. Research question

Our project aims to provide clarity to mainstream media sources about where they stand in relation to their audience. Does their sentiment about a topic match that of the average twitter user? Or are their opinions alienating consumers of their media. Where does a mainstream media outlet rank with respect to twitter users? Does any one outlet have an upper hand?

We also aim to provide recommendations of which media outlets a twitter user should follow based on the outlets that other twitter users agree with the most. For example, in our analysis we found that USA Today had the highest rank. Therefore It would make sense to recommend USA Today to an average twitter user. For a more customized option, we have also performed analysis of which outlet twitter users agree with the most based on a particular topic. So a twitter user making a lot of tweets about the “vaccine mandate” would be recommended to Reuters World as the majority of twitter users shared a very similar sentiment polarity with that particular outlet.

4. Methodology

Data Source:

We used a google-news API to scrape the latest news articles from various publishers and by extension the latest news topics. Using the tweepy API, we gathered the tweets made by both mainstream media outlets and regular twitter users.

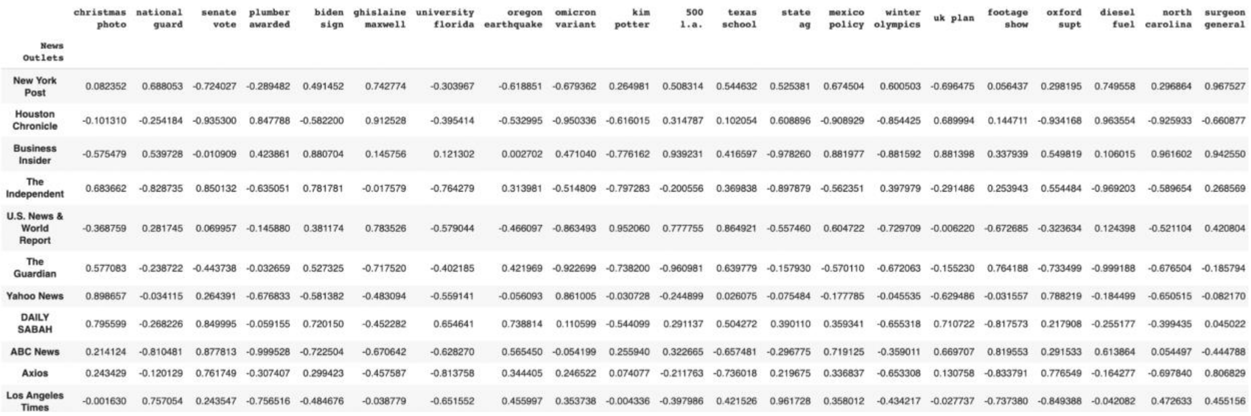
https://rapidapi.com/newscatcher-api-newscatcher-api-default/api/google-news

Data Preprocessing:

For our data preprocessing, we used the google news API to scrape news articles and all information related to that article such as the title/headline and the publisher. From these news headlines we extracted the topic in the form of a bigram by using tokenization to remove stopwords and punctuations. We stored these topic bigrams in a topics list. I extracted the publishers of the articles and stored them in a list of publishers. Then gathered their respective twitter handles. Using these twitter handles, we were able to pull tweets from their timelines containing those particular bigrams and stored those filtered tweets in a list. As for our feature selection, we found that the API would update twice a day with new stories. Therefore, we decided to move forward with the 30 most recent topics so as to provide analysis of the freshest topics. This also helped us find relevant tweets on the twitter timelines of the mainstream media outlets.

Data Processing:

After preprocessing our data, we imported the TextBlob library for NLP Sentiment Analysis. For the filtered tweets from each twitter handle, we then applied the sentiment analyzer and stored the sentiment values in a sentiment polarities list. We then took the average of the sentiment polarities list and added the average value into another list which became the value in a dictionary with the topic as the key (eg. {“christmas photo”: [0.082352, -0.101310,...], “national guard”:[...]}). We use this dictionary to populate our dataframe. In the end our dataframe looked something like this:

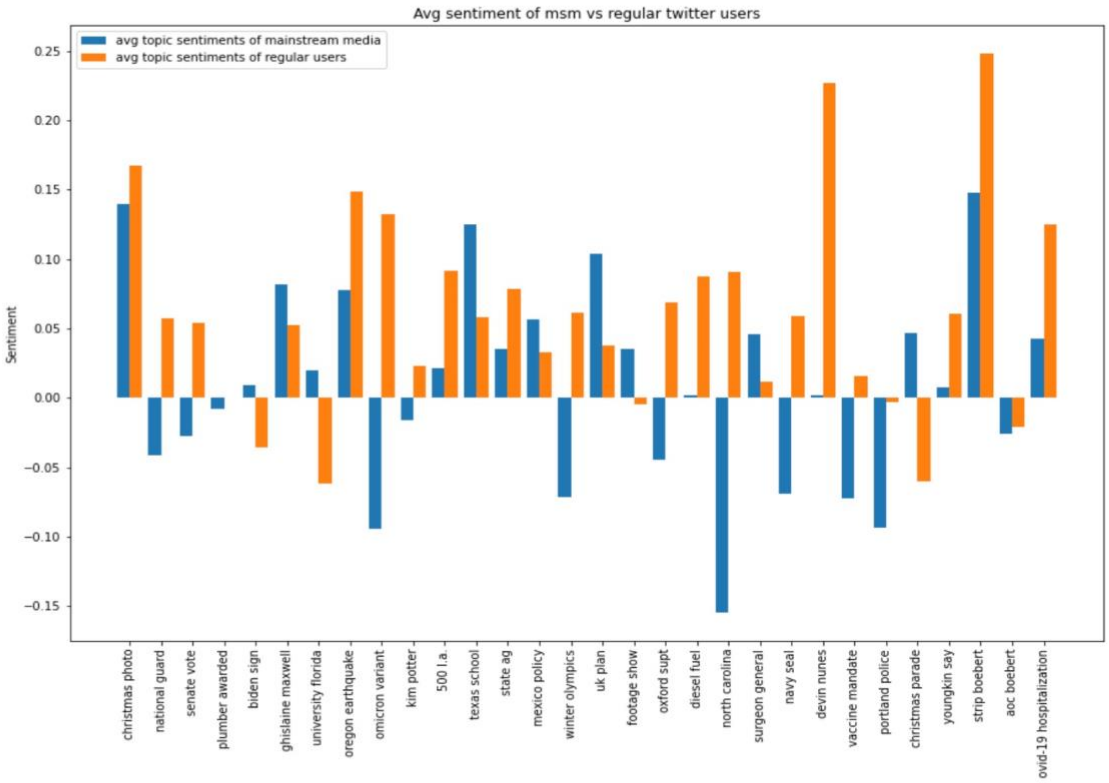


News Outlets	christmas photo	national guard	senate vote	plumber awarded	biden sign	ghislaine maxwell	university florida	oregon earthquake	omicron variant	kim potter	500 l.a.	texas school	state ag	mexico policy	winter olympics	uk plan	footage show	oxford supt	diesel fuel	north carolina	surgeon general
New York Post	0.082352	0.688053	-0.724027	-0.289482	0.491452	0.742774	-0.303967	-0.618851	-0.679362	0.264981	0.508314	0.544632	0.525381	0.674504	0.600503	-0.696475	0.056437	0.298195	0.749558	0.296864	0.967527
Houston Chronicle	-0.101310	-0.254184	-0.935300	0.847788	-0.582200	0.912528	-0.395414	-0.532995	-0.950336	-0.616015	0.314787	0.102054	0.608896	-0.908929	-0.854425	0.689994	0.144711	-0.934168	0.963554	-0.925933	-0.660877
Business Insider	-0.575479	0.539728	-0.010909	0.423861	0.880704	0.145756	0.121302	0.002702	0.471040	-0.776162	0.939231	0.416597	-0.978260	0.881977	-0.881592	0.881398	0.337939	0.549819	0.106015	0.961602	0.942550
The Independent	0.683662	-0.828735	0.850132	-0.635051	0.781781	-0.017579	-0.764279	0.313981	-0.514809	-0.797283	-0.200556	0.369838	-0.897879	-0.562351	0.397979	-0.291486	0.253943	0.554484	-0.969203	-0.589654	0.268569
U.S. News & World Report	-0.368759	0.281745	0.069957	-0.145880	0.381174	0.783526	-0.579044	-0.466097	-0.863493	0.952060	0.777755	0.864921	-0.557460	0.604722	-0.729709	-0.006220	-0.672685	-0.323634	0.124398	-0.521104	0.420804
The Guardian	0.577083	-0.238722	-0.443738	-0.032659	0.527325	-0.717520	-0.402185	0.421969	-0.922699	-0.738200	-0.960981	0.639779	-0.157930	-0.570110	-0.672063	-0.155230	0.764188	-0.733499	-0.999188	-0.676504	-0.185794
Yahoo News	0.898657	-0.034115	0.264391	-0.676833	-0.581382	-0.483094	-0.559141	-0.056093	0.861005	-0.030728	-0.244899	0.026075	-0.075484	-0.177785	-0.045535	-0.629486	-0.031557	0.788219	-0.184499	-0.650515	-0.082170
DAILY SABAH	0.795599	-0.268226	0.849995	-0.059155	0.720150	-0.452282	0.654641	0.738814	0.110599	-0.544099	0.291137	0.504272	0.390110	0.359341	-0.655318	0.710722	-0.817573	0.217908	-0.255177	-0.399435	0.045022
ABC News	0.214124	-0.810481	0.877813	-0.999528	-0.722504	-0.670642	-0.628270	0.565450	-0.054199	0.255940	0.322665	-0.657481	-0.296775	0.719125	-0.359011	0.669707	0.819553	0.291533	0.613864	0.054497	-0.444788
Axios	0.243429	-0.120129	0.761749	-0.307407	0.299423	-0.457587	-0.813758	0.344405	0.246522	0.074077	-0.211763	-0.736018	0.219675	0.336837	-0.653308	0.130758	-0.833791	0.776549	-0.164277	-0.697840	0.806829
Los Angeles Times	-0.001630	0.757054	0.243547	-0.756516	-0.484676	-0.038779	-0.651552	0.455997	0.353738	-0.004336	-0.397986	0.421526	0.961728	0.358012	-0.434217	-0.027737	-0.737380	-0.849388	-0.042082	0.472633	0.455156

You can see the news outlets as the rows and each column is a topic bigram. Each and every value is the average sentiment of a topic with respect to the mainstream media outlet.

We then performed various EDA to gain some insight into our data. We visually compared the sentiments of both mainstream media outlets and average tweeters with respect to various topics. We also visualized the rank of each mainstream outlet and recommended the outlet which is the most popular with respect to a particular topic as well as recommending the highest ranking outlet.

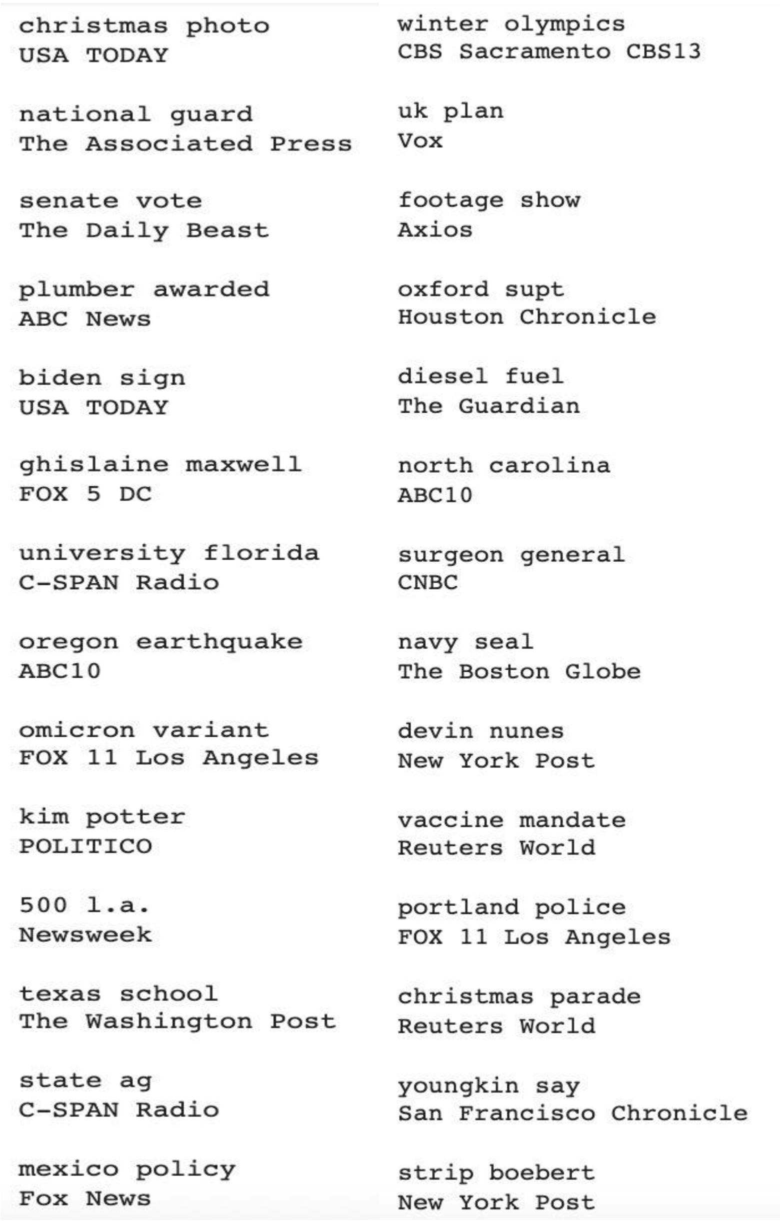
<u>Visualization results:</u>



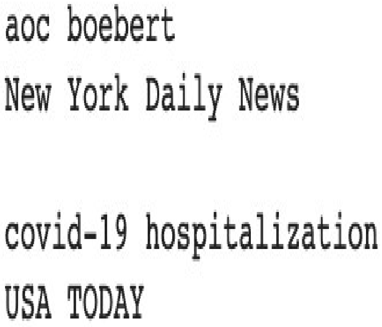
The bar graph above compares the average sentiment of twitter users about a topic with the average sentiment of mainstream media. As we can see, there exist disparities between the mainstream sentiment and the sentiment among twitter users. For example, the topic about the winter Olympics. Mainstream media has been talking about it often with a negative sentiment due to China’s response to covid-19. Whereas regular twitter users don’t really care about the politics and just enjoy watching the athletes. Therefore there is a huge discrepancy between the sentiments of mainstream media and regular twitter users.

Reference link:

https://www.google.com/amp/s/www.bbc.com/news/uk-59582137.amp

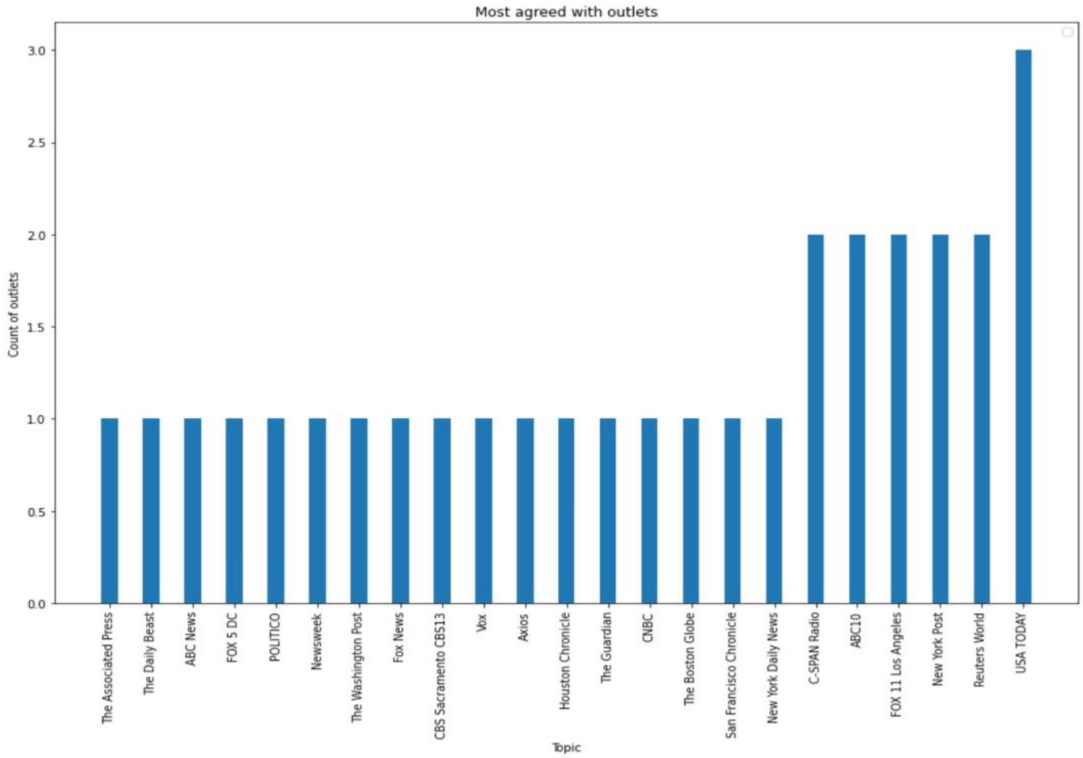


```
christmas photo              winter olympics
USA TODAY                    CBS Sacramento CBS13

national guard               uk plan
The Associated Press         Vox

senate vote                  footage show
The Daily Beast              Axios

plumber awarded              oxford supt
ABC News                     Houston Chronicle

biden sign                   diesel fuel
USA TODAY                    The Guardian

ghislaine maxwell            north carolina
FOX 5 DC                     ABC10

university florida           surgeon general
C-SPAN Radio                 CNBC

oregon earthquake            navy seal
ABC10                        The Boston Globe

omicron variant              devin nunes
FOX 11 Los Angeles           New York Post

kim potter                   vaccine mandate
POLITICO                     Reuters World

500 l.a.                     portland police
Newsweek                     FOX 11 Los Angeles

texas school                 christmas parade
The Washington Post          Reuters World

state ag                     youngkin say
C-SPAN Radio                 San Francisco Chronicle

mexico policy                strip boebert
Fox News                     New York Post
```

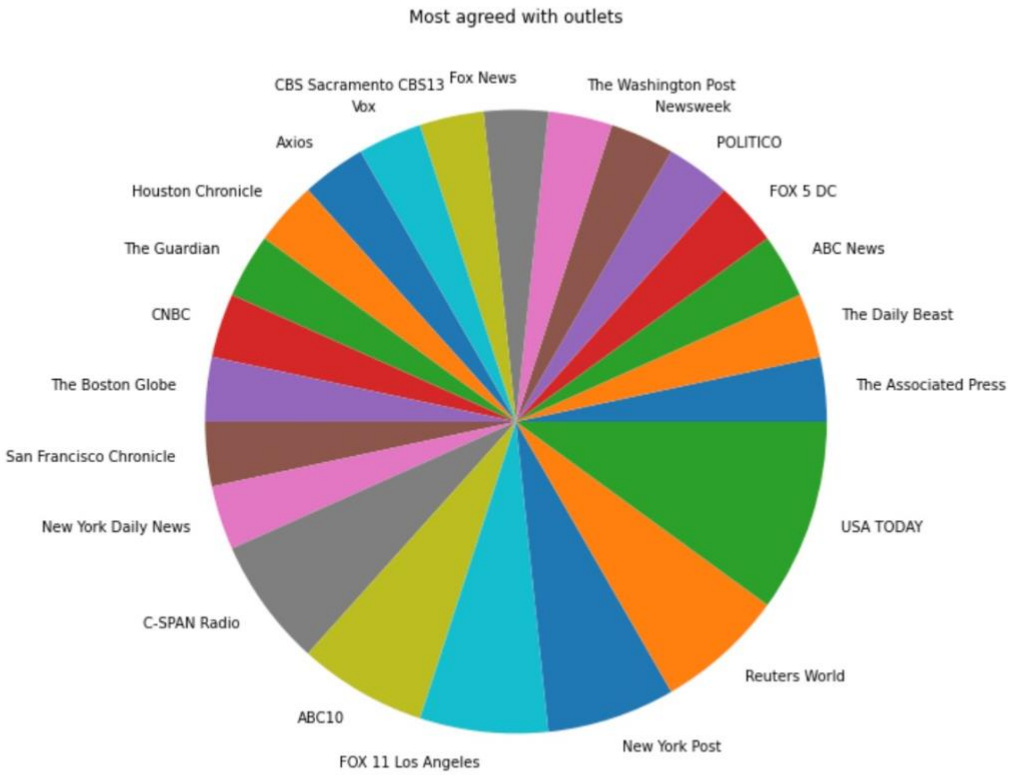


```
aoc boebert
New York Daily News

covid-19 hospitalization
USA TODAY
```

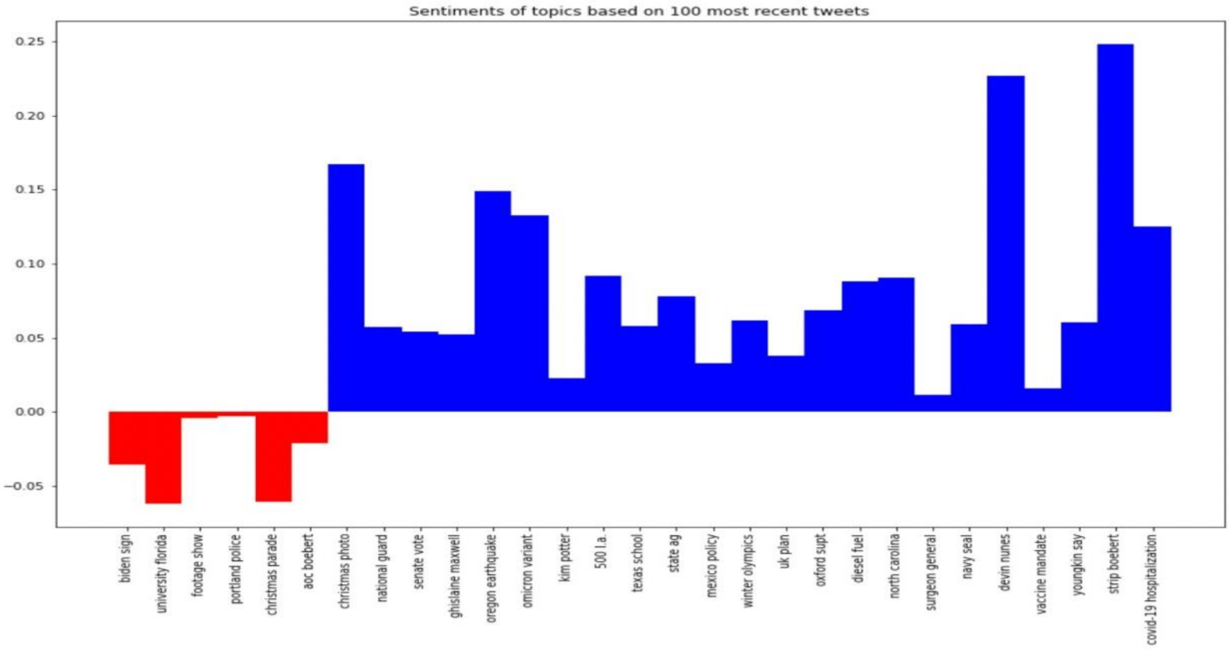
In the printed data above, we can see that underneath each topic name is the name of a media outlet. This visualization shows which media outlet people agreed with the most upon regarding a particular topic. For example, we can see that with respect to the topic “Mexico policy”, people agreed with Fox news more than any other outlet. Another example is the topic of “State AG” where people agreed with C-SPAN Radio the most. We were able to obtain this visualization by subtracting the avg sentiment of that topic with respect to the media outlet and the avg sentiment of twitter users. Whichever outlet had the least difference was the one that average twitter users agreed with the most.



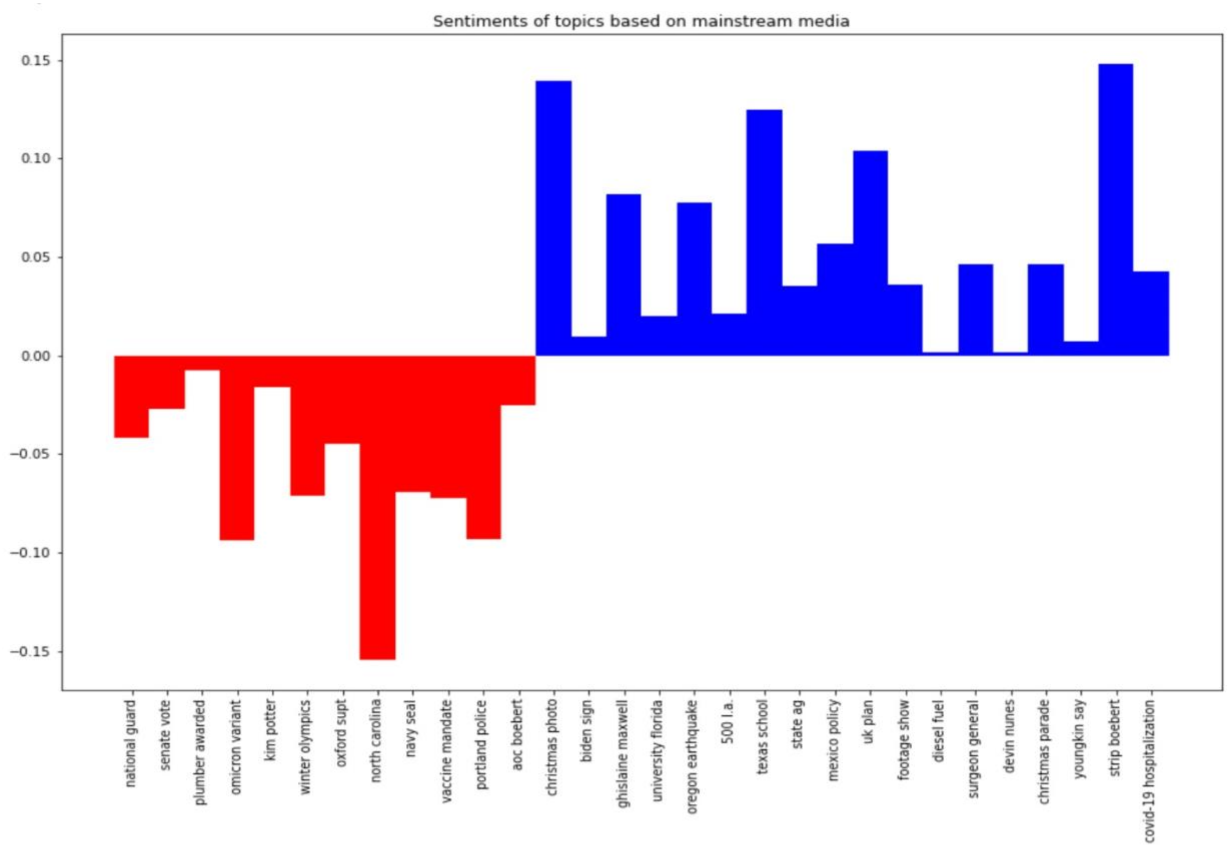
This visualization shows the general rank of the mainstream media outlets in ascending order. This is with respect to popularity and integrity. The measure we used for integrity was the number of times average twitter users agreed with a mainstream outlet. For example, twitter agreed with USA Today 3 times whereas other outlets were agreed upon only 1-2 times. There were some times when no outlet was ranked higher than another. But in this instance, we found USA Today to rank the highest followed by Reuters World, New York Post, Fox 11 LA, and ABC 10.



Above is a pie chart visualization of the most agreed upon outlets. As we can see, USA Today displays the biggest chunk as it was the outlet people agreed with the most often.



The visualization above depicts a bar graph of the average sentiments of topics based on the 100 most recent tweets made about the topic, i.e. the sentiments of average twitter users.



The visualization above depicts a bar graph of the sentiments of mainstream media with respect to particular topics.

5. Discussion

Why does your methodology work (or does not work)?

The methodology we implemented works. Although it is slow to run, our goal to rank outlets works. In our data, we found that USA Today is the outlet that people agreed with most often. There are occasions when there is no ranking because there is no single outlet which has the highest rank. No one outlet is being agreed upon more than the other.

Why are your findings meaningful?

Our project ranks media sources by the number of times people agreed the most with their sentiment about a topic. For example, people might have agreed the most with CNN on 7 topics while they agreed with the New York Times only 4 times. So we can say that CNN ranks above the New York Times. This tells an msm outlet where they rank with respect to their competitors. Thereby, telling them to improvise their strategies and presenting news that resonates with people. This project will try and force msm outlets to represent the average person and provide news which is more factual and relevant to the average person. As msm outlets compete to rank the highest and match the sentiments of regular people, their news stories will also become more relatable to the average joe. This also increases the viewership and followers of the msm outlet. It’s a win-win situation!

What are the limitations and how to improve?

Some mainstream outlets don't have twitter accounts. Twitter does not have a way to search for a user’s twitter handle with the screen name. There is a limitation to the number of tweets obtained. Also could use more news API’s to bolster the amount of media sources and topics acquired. Sometimes bigrams are not good enough and it may be required to use trigrams. Limitations to when twitter users use sarcasm and jokes as sentiment analysis is then skewed. We also would need to make the populating of mean polarities faster. For populating a (48,121) dataframe, it takes 5.5 hours to get the data back from the API, process it, compute the average polarity for all the columns, and populate the dataframe. The biggest limitation in our project was the limited capabilities of the tweepy API.

References

https://computationalsocialnetworks.springeropen.com/articles/10.1186/s40649-019-0063-4

https://www.analyticsvidhya.com/blog/2021/06/twitter-sentiment-analysis-a-nlp-use-case-for-beginners/

https://towardsdatascience.com/step-by-step-twitter-sentiment-analysis-in-python-d6f650ade58d

https://www.nature.com/articles/s41598-021-86510-w

https://www-cs.stanford.edu/people/alecmgo/papers/TwitterDistantSupervision09.pdf

http://www.lrec-conf.org/proceedings/lrec2010/pdf/385_Paper.pdf

https://aclanthology.org/C10-2005.pdf

https://www.google.com/amp/s/www.bbc.com/news/uk-59582137.amp

https://textblob.readthedocs.io/en/dev/quickstart.html