**CSE 5095: Social Media Mining & Analysis**

**Fall 2020: Project Descriptions**

**Disclaimers:**

**This raw data was collected during the Covid-19 pandemic and BLM protests, over the past 6 months. It represents people’s reactions and opinions to important events and occurrences during the pandemic and protests. The data may contain offensive and derogatory language, sometimes even targeted towards specific groups. Please note that I do not endorse the use of such language towards anyone in this class. Moreover, I do not specifically support or oppose any point of view represented in the data. We will use this data solely for academic projects, to demonstrate that social media data is rich with people’s attitudes and opinions, and it can be mined to extract such opinions.**

**Notes:**

1. **You can access the raw data at:**

[**https://www.dropbox.com/sh/q6r2571o0kc4wqt/AADZv3GjqiB0zS9zU2XZfng4a?dl=0**](https://www.dropbox.com/sh/q6r2571o0kc4wqt/AADZv3GjqiB0zS9zU2XZfng4a?dl=0)

1. **I will be updating and adding labeled data continuously during the course of the semester.**
2. **In your projects, please use only the labeled data points.**

**Projects 1, 2 Attitude about Vaccines**

The coronavirus pandemic has fueled anti-vaccine protests in closed Facebook groups, as well as on public social media platforms platforms. Anti-vaxxers, already skeptical and suspicious about the ill-effects of a vaccine are questioning the safety of a rushed vaccine. This has made public health experts nervous about the widespread adoption of a vaccine, which is necessary to build herd immunity. They recognize the need to control the messaging, and to dispel the myths and lies propagated by anti-vaxxers for a buy-in of the vaccine. The central tenets of the anti-vaccine movement are the vaccines are unsafe (they cause autism, disabilities, and may even be fatal), they represent a monetary alliance between the federal government and big pharmaceutical companies. They also harbor a deep mistrust of the Deep State/political agenda/media that seeks to interfere in individual’s lives and liberties.

Two data sets were collected one on May 7, and July 17, using the crawling seed #vaccines. Each tweet is labeled as ‘P’ and ‘A’, where ‘P’ refers to pro-vaxx opinion, and ‘A’ refers to anti-vaxx opinion. Note that anti-vaxx opinion also includes those against the Covid-19 vaccine in particular, although they may not be against vaccinations in general. Any neutral tweet, or a tweet that does not have clear anti-vaxx leaning is also classified as pro-vaxx, unless there are some clear indications that suggest otherwise. Such clues may come in the form of hashtags, Twitter handles.

There will be two projects based on these two data sets, one data set per team. The idea is to understand if there are any longitudinal/evolving trends in people’s outlook on the Covid-19 vaccine, as the pandemic has raged on, and continued to cause widespread disruption. It is especially curious to know if the public stance on the vaccine softened, given that by July the coronavirus was spreading like wildfire in the South and the West, and states in these regions had to roll back their reopening. Moreover, to prevent the spread of discordant information, we need to identify it. For this purpose, we need to have trained machine learning models that can be deployed in real time. To build this capability, we will explore the feasibility of using the labeled May 7 data to predict the anti- and pro-vaxx tweets from the July 17 data, and vice versa. Thus, these two data sets will make for two separate, yet loosely coordinated projects. Such coordination will occur in the latter part of the semester. For now, the two groups will proceed independently with their projects.

**Project #1: vaccinesMay7**

An example of pro-vaxx and anti-vaxx tweets from the May 7 data set are shown below

**Anti-vaxx (‘A’)**

*Sleep paralysis but instead of a demon in the corner of your room itâ€™s bill gates injecting you with autism filled vaccines while wearing a hazmat suit to protect himself from 5g radiation*

**Pro-vaxx (‘P’)**

*@fox4kc @lindawagarkc Vaccines save lives.*

**Project #2: vaccinesJuly19**

An example of pro-vaxx and anti-vaxx tweets from the May 7 data set are shown below

**Anti-vaxx (‘A’)**

*Walmart, Kroger, Aldi's, Publix...these tirant store CEO'S making you wear a mask to enter...People-stand up! Don't give into them. The next thing will be mandatory vaccines, then, what? Cashless-microchip.*

**Pro-vaxx (‘P’)**

*Sars-Cov-2 virus is here to stay. We will see surges as it expands around the globe, until it becomes endemic like influenza with seasonal surges. Lockdowns only delay the surges. Effective vaccines will be the way to control illness, but the virus will not be eradicated.*

**References**

1. <https://www.washingtonpost.com/national/trumps-promise-of-warp-speed-fuels-anti-vaccine-movement-in-fertile-corners-of-the-web/2020/05/20/c2b3d408-9ab2-11ea-89fd-28fb313d1886_story.html>
2. <https://www.adn.com/nation-world/2020/05/21/trumps-promise-of-warp-speed-vaccine-development-fuels-anti-vax-movement-online/>
3. <https://www.msn.com/en-us/news/us/us-coronavirus-america-broke-its-single-day-record-of-new-cases-at-least-9-times-in-a-month/ar-BB16Vdjp?ocid=spartan-dhp-feeds>

**Project 3,4: Debate Over Masks**

Mask wearing has become a topic of fierce debate in the United States. People opposed to mask mandates have staged protests, and one local health official had to even quit her job after receiving a death threat for a mask order. It's understandable if some people remain skeptical, since, at the beginning of the pandemic, public health officials in the U.S. said the general public didn't need masks.

Proponents of masks point to several studies that recommend their use to slow the spread of Covid. However, most of the studies have looked at face mask use in health care, not community settings. And they were observational, not the gold standard of science, a randomized controlled trial. The U-turn regarding masks and the subsequent political divide over them has come to symbolize the chaos of the U.S. response to the still-raging pandemic. Some counties, where the coronavirus has surged out of control has imposed mask mandates, which has further outraged their residents.

Data was collected twice, one month apart, using the crawling seeds *#wearadamnmask, #nomaskforme, #maskupamerica, #masksareforsheep, #nomasks, #nomasksforme, #nomaskmandate, #masksareforsheep, #antimaskers, #maskitorcasket* in July 2020 and August 2020. Masks came into sharp spotlight in this one-month period because of the tussle surrounding the reopening of schools, and students returning to college campuses. Masks also became a hot button issue during this period when Democratic Presidential Candidate Joe Biden suggested that if elected he will issue a national mask mandate. In the same period, leading public health experts, including the CDC promoted the use of masks as “life-saving”, highlighting that if everyone committed to wearing masks, we could save a significant number of American lives. The two data collection epochs one month apart, which was eventful, was intended to assess if there were any shifts in the attitude due to the spotlight on masks, and their effectiveness. These two data sets will make up for two individual, yet coordinated projects, similar to the project on vaccines. The idea is that similar to anti-vaxxers, anti-maskers also spread discordant information which needs to be curbed by employing the classifiers in real-time. Because the tweets supporting and opposing masks did not align according to the hashtags, we labeled the data into for and against groups. Representative tweets from each category are included below:

**Project #3 – July 2020**

**Tweets Supporting Masks (‘Y’)**

*Erry time I ride @trimet to work thereâ€™s always a couple ppl not wearing masks... like really? #WearADamnMask*

**Tweets Opposing Masks (‘N’)**

*History REPEATING itself #NoMasks #NoMaskOnMe #BLMIsADomesticTerroristGroup* [*https://t.co/oJjoPA9WW9*](https://t.co/oJjoPA9WW9)

**Project #4 – August 2020**

This data also contained a number of public safety announcements (PSAs) from schools, colleges and sports teams. There were tweets that expressed political opinion regarding the conventions, wildfires in California, and the BLM protests without the express mention of masks other than the hashtag. These data were eliminated to build a high quality data set that truly reflects the public opinion about masks instead of other peripheral, political issues. Examples of tweets supporting and opposing masks are:

**Tweets Supporting Masks (‘Y’)**

*At least my mask hides my pimple #moreimportantlyissaveslives #WearADamnMask* [*https://t.co/hbxquonf8R*](https://t.co/hbxquonf8R)

**Tweets Opposing Masks (‘N’)**

*@healthvermont @CDCgov We are tired of the government telling us to stay safe. Freedom Trumps safety in America. Fellow Vermonters tell the government, â€œWe will not comply!â€ #NoMasks #freedom #Vermont*

**References**

1. <https://www.npr.org/sections/health-shots/2020/06/21/880832213/yes-wearing-masks-helps-heres-why>
2. <https://www.vox.com/2020/8/21/21395570/biden-dnc-national-face-mask-mandate>
3. <https://www.msn.com/en-us/news/us/mask-mandates-continue-to-spark-outrage-as-virus-vaccine-trials-ramp-up/ar-BB17fJeO?ocid=spartan-dhp-feeds> <https://www.washingtonpost.com/world/2020/06/25/face-masks-america-divided/>

**Project 5-9: Anti-Lockdown Protests**

Anti-government sentiment percolated among far-right extremists in weeks after the stay-at-home orders were issued by the governors to prevent the spread of the coronavirus. Trump’s “LIBERATE” tweets further fueled these sentiment, because extremists viewed this as a call to action, and extremist groups wondered whether Trump was advocating for armed conflict, an event they termed as “Boogaloo”. Reopen protests were organized in many states, with North Carolina being the first. Michigan, Pennsylvania, Virginia, Colorado, Texas, California, New York, New Jersey, Wisconsin followed. Organizers of these protests sported affiliation with hate groups (Proud Boys, Boogaloo, etc) who are advocates of gun rights and supporters of the 2nd Amendment. Boogaloo, also known as Boogaloo Bois or Bois is a loosely organized far-right militia movement that seeks to spark a second civil war in the United States. It is a loose association of gun owners who mostly lean right-libertarian or anarchist. For example, the husband of Reopen NC cofounder Ashley Smith saying supporters must be "willing to kill people."

The protests that received the most media coverage and notoriety were from North Carolina, Pennsylvania, Michigan, and California. These states also represent a continuum on the political spectrum. North Carolina is heavily Republican, California is Democratic, and Michigan and Pennsylvania are battlegrounds. Corresponding state-specific hashtags appeared and trended on Twitter. The protestors made their discontent vocal, and received significant attention from the media. Tweets supporting the protests criticized the stay at home orders because they consider it as a power grab and tyranny, and are mindful of their adverse economic impact. They contend that these orders are unconstitutional because they take away people’s liberties and freedoms, and some threaten ballot recalls. These tweets may also insult democrats, and their liberal policies. On the other hand, tweets against the protests and in favor of continuing the lockdown refer to the impact of these protests on increasing the number of Covid cases and deaths. They refer to the protestors as terrorists, expressing horror at the fatalities, etc. The media and other public health experts also criticized the protestors for flouting social distancing and mask rules. What was not heard was the voice of ordinary people who stayed in the comfort and safety of their homes, following the rules laid by their respective governors. Data was collected from Twitter for these protests, and will be analyzed to assess whether it is possible to distinguish between pro-protest and anti-protest groups. Because of their loose affinity with Boogaloo anarchists, it is important to be able to identify tweets that are sympathetic to the protests. This is especially important when the platforms themselves do not seem to be doing enough to delete or tag such posts.

**Project 5: Michigan, Operation Gridlock (April 21)**

**Supporting Protests**

*This is what you get when you vote Democrat, Michigan! STREETS PACKED IN LANSING -- Hundreds More on Their Way to Protest Tyrannical Governor -- Citizens Protest Crazed Governor Whitmer's Police State Policies (VIDEO) https://t.co/AgX645OptP*

**Opposing Protests**

*All I can say is Gob Bless those patriots who grabbed their AK-47, filled up their Dodge Ram with their stimulus check and headed to Lansing to ignore the CDCâ€™s social distancing guidelines! These are the REAL heroes!*

**Project 6: Michigan (May 1)**

**Supporting Protests (Y)**:

*@t42592 It's still not too late to put #BALLOTRECALLGOVERNORS on the ballot! #MichiganProtest*

**Opposing Protests**:

*In 24 hours 2,390 American's died from Covid-19, 30,000 NEW cases yesterday and as of an hour ago 63,500 have died in the last 8 weeks. Do Michigan protesters care? No. When was the last time you ever saw..*

**Project 7: North Carolina**

**Supporting Protests (Y)**

*@RoyCooperNC stop telling us how to live our lives. We have learned a lot about this virus and it's time for you to release your money hungry hold and dictatorship on us. @POTUS @Mike\_Pence #YoureNotMyRealDad #ReopenNC*

**Opposing Protests (N)**

*To everyone making cracks about North Carolina, rest assured that the VAST MAJORITY of North Carolinians are NOT as stupid and ignorant as the members of #ReopenNC.*

**Project 8: Pennsylvania**

**Supporting Protests (Y)**

*Youâ€™re killing us @GovernorTomWolf !!!Thank you for standing with Pennsylvania @SenMastriano and @realDonaldTrump #ReOpenPA #ImpeachGovernorWolf https://t.co/HyFInQmJwd*

**Against Protests (N)**

*Iâ€™ve never been more embarrassed to call myself a Pennsylvanian. Do not #ReOpenPA for the sake of our health. Wouldnâ€™t be shocked if our case numbers skyrocket in respond of this idiotic protest.*

**Project 9: California**

**Supporting Protests (Y)**

*We now have testing available for all who are symptomatic, an FDA approved treatment, and an anecdotally proven treatment (HCQ/Zpack/Zinc). The curve is sloping DOWN! Time to #OpenCalifornia #OpenAmerica*

**Opposing Protests (Y)**

*Until you can provide me with links from REPUTABLE SCIENTISTS AND DOCTORS that says that Coronavirus is all BS then STOP preaching on how we should open back up. CA had the highest recorded number of deaths in over a week bc people are getting LAZY. #DoNotOpenCalifornia*

We may also have a chance to corroborate these opinions with some online polls and surveys in the references.

**References**

1. <https://indyweek.com/news/northcarolina/unc-wilmington-professor-mike-adams-raleigh-police-reopen-nc/>
2. <https://www.csmonitor.com/USA/Politics/2020/0504/Guns-in-Michigan-Capitol-Defense-of-liberty-or-intimidation>
3. <https://www.jacobinmag.com/2020/04/coronavirus-pandemic-lockdown-protests-ubi>
4. <https://gandernewsroom.com/2020/07/10/michigan-supports-whitmer-pandemic-leadership/>
5. <https://www.snopes.com/news/2020/05/20/michigan-lockdown-protesters/>
6. <https://www.businessinsider.com/michiganders-approve-of-whitmer-on-coronavirus-despite-protests-poll-2020-4>
7. <https://changeresearch.com/post/pollercoaster-2020-michigan-voters-on-covid-and-trump-biden/>
8. <https://www.washingtonpost.com/politics/2020/08/10/movement-against-coronavirus-lockdowns-is-still-going-still-angry/>
9. <https://www.abc27.com/news/us-world/national/virus-restrictions-fuel-anti-government-boogaloo-movement/>
10. <https://www.nbcnews.com/tech/security/trump-s-liberate-tweets-extremists-see-call-arms-n1186561?fbclid=IwAR1iRdKWk3-jfqdWNSqGIDBkGefFinkHujOuodHeAHjxLnS5eR23KlhiWzY>
11. <https://www.rawstory.com/2020/05/husband-reopen-nc-leader-willing-to-kill-people-in-resistance-to-emergency-orders/>
12. <https://arxiv.org/ftp/arxiv/papers/2008/2008.09440.pdf>
13. <https://www.voanews.com/covid-19-pandemic/virus-restrictions-fuel-anti-government-boogaloo-movement>
14. <https://www.facebook.com/Fakertarians-Is-A-Communist-Group-466829853887850/>
15. <http://ncccommuter.org/boogaloo-where-ar-15-meets-aloha/>
16. <https://www.reuters.com/article/us-zuckerberg-kenosha/zuckerberg-says-facebooks-failure-to-remove-militia-page-an-operational-mistake-idUSKBN25O2ZJ>

**Project 10, 11. Michigan Protests – Offensive/Non Offensive**

I observed that tweets on both sides contained a significant amount of offensive and toxic content, especially, in the Michigan data sets. Beyond the definitions in the dictionary of these terms, there is generally no consensus or uniformity on what constitutes offensive or hateful speech. Labeling hate speech is difficult because of different types of labels that can confuse the labelers. e.g., offensive language, hate speech, aggressive behavior, cyberbullying, etc. Many different types of labels are used hate, offensive, abusive, racism, sexism, harassment. In this labeling, I have looked for comments that may insult, humiliate, bully, incite anger and even promote violent behavior. These insults and humiliation could be based on immutable or inherent characteristics such as race, gender, skin color, nationality, etc. In this form, it is considered to be hate speech. This is the most egregious form. Speech that hurls insults because of intellect, political opinion, affiliation and views, cultural and geographic stereotypes is also intended to harass and bully. For example, in harassment based on appearance, all dimensions of physical appearance such as hair style or looks are considered fair game for ridicule. Fat shaming [1] and body shaming are critical sub-types. In harassment based on cultural and geographic stereotypes, it may be common to refer to people as rednecks and white trash. Referring to people as animals such as rats and snake is also derogatory and humiliating. Tweets may also express hostility or animosity towards a particular group. The distinction between hostility and criticism is critical to protect freedom of speech. Sometimes a word is added for its alternate meaning or intention or for emphasis, as an instance the f\*\* word might be used to emphasize ruin or damage. For example, in *“If this virus has taught me anything it’s that Americans literally have zero fucking clue what the constitution actually says. #COVID #MichiganProtest”*, the offensive word is added for emphasis. On the other hand, in *“Well there were plenty of fat f\*cks at the protest!”*, the word indicates appearance-based harassment. We also include directed harassment such as an @mention that says "you are a loser and I hope you die". Because there are different ways of labeling toxic content (hate, abusive, offensive, harassment, hostile), we use an all -inclusive category of toxic content. Words or phrases insinuating mental illness are also considered offensive such as nuts, numbnuts, etc.

Finally, it is important to note that the mere presence of a profane word does not indicate abusive or bullying intent towards a specific person or a group. For example, “*Thought racism was thinking that one "skin color" was more superior than another, so how is people of ALL skin colors wanting to go back to work racist? #MichiganProtest* [*https://t.co/sgr7Q1D9Sf*](https://t.co/sgr7Q1D9Sf)*.”* Here the use of the term racist does not constitute harassment or as being offensive. Thus, the detection of offensive speech is context dependent, and cannot be based on simple words. It is important to capture the intent of the tweet in the classification; it cannot be based on simply on the use of specific words. Although the protests were organized by right-wing extremist groups affiliated with Boogaloo, the use of offensive language was found on both sides. Offensive tweets are labeled as ‘O’, and non-offensive tweets are labeled as ‘NO’. The two data sets, one from April 21 and the other from May 1 constitute two independent projects initially. Eventually, we will use one data set for training and the other for testing.

**Project #10: April 21**

**Offensive Tweets**

*i just saw a tweet where someone thought they exposed the Deep State by parking their car in Lansing yesterday. what a fucking mung.*

**Non Offensive Tweets**

*I find that while people have a right to peaceful protest, people protesting outside the capital in Lansing are irresponsible and selfish. I want to go back to work, too, but not at the risk of removing myself permanently from the lives of my family. #upwith13*

**Project #11: May 1**

**Offensive Tweets**

*On our way to 100k dead and Mr. #maga only thinks about his #kkk rallies. Must have missed the #Incel Pride gathering of #ammosexuals &amp; their toys at the #reopen #MichiganProtest. Donâ€™t worry <f0><U+009F><U+008D><U+008A><f0><U+009F><U+0092><U+00A9><f0><U+009F><U+00A4><U+00A1> all the methheads and deathcultists are still out spreading your racists hate. https://t.co/8ftGqFc8V6*

**Non Offensive Tweets**

*No Gretchen, you and #HillaryClinton destroyed your political party, that's all you did. You could have done it differently. #MichiganProtest Whitmer says #Michigan protests 'depicted some of the worst racism and awful parts of our history' https://t.co/zzkpfvijqr*

What is exciting about this project is that under ordinary, normal circumstances, there is very little offensive content on social media. As a result, it is not very well-known how such toxic content can be identified accurately from social media feeds. These, however, are not ordinary times, and this data, is rife with such toxic content presenting an opportunity to detect offensive speech in a real-life situation. Hate breeds hate, it angers and incites, and makes a compromise difficult to achieve. Therefore, regardless of the ideology or the philosophy it is never acceptable to use offensive and toxic rhetoric that harasses and insults, and it is important to detect such content.

**References**

1.Facebook community standards. <https://www.facebook.com/communitystandards/hate_speech>.

2. Content Moderation & Hate Speech. Should Platforms like Facebook do More? <https://www.msn.com/en-us/money/companies/we-re-in-we-re-out-the-north-face-becomes-the-first-major-company-to-boycott-facebook-as-the-calls-for-advertisers-to-walk-out-of-the-platform-in-july-intensify/ar-BB15J74T?ocid=spartan-dhp-feeds>

3. The projects directory also has a Harrasment-Corpus.

**Project #12, 13: Proud Boys**

On August 22, far-right activists including Proud Boys, exchanged blows with Antifa and BLM protestors in Portland, as Portland Police just stood by. Proud Boys has been designated as a “hate group” by the Southern Poverty Law Center. Following the two-hour brawl which qualified to be a riot, the hashtag “Proud Boys” was trending on Twitter. Supporters of Proud Boys noted that they were engaged in similar violent tactics as BLM protestors. Detractors expressed horror that the police just stood by, speculating that at least some of the police were either a part of the Proud Boys group, or on the side of proud boys. Sympathizers also retorted that whatever the police had been doing with Antifa protestors (step back and watch) is exactly what they did with the Proud Boys. Portland police declared the assembly unlawful and to be a riot after the Proud Boys left. Another narrative circulating was that it was a staged act by Antifa/BLM protestors to frame the Proud Boys. According to them, in reality, it was just Antifa vs. Antifa. A similar violent protest also broke out in Kalamazoo, Michigan on August 16, with similar narrative

Because Proud Boys is a hate group, the first goal is to identify tweets that are sympathetic to this hate group. They may or may not support the ideological underpinning of Proud Boys, but they may support this particular conflict that they have gotten involved in. Tweets that “educate” about what Proud Boys are, or dispute that there were proud boys are also labeled as being supportive to their mission. A number of tweets are directed at two conservative political commentators Mr. Andy Ngo, and Ms. Savannah Herandez. Both supporting and opposing tweets were directed at these commentators. Tweets that express belief that these protests are staged are also labeled as being supportive of the Proud Boys, especially because they believe that these staged protests are funded by liberals such as George Soros, for fake propaganda. Secondly, we also want to detect offensive and toxic content in the discourse. Although Proud Boys is labeled as a hate group (they have appeared alongside other hate groups such as “Unite the Right” rally in Charlottesville), generally, hateful and offensive content against anyone should be avoided because it can incite violence.

Each tweet has three labels Y/N (supporting or opposing proud boys), and O/NO (offensive/non offensive) and S/NS (Sarcastic/Non Sarcastic). For now, we will focus on the first two classifications. (Only offensive tweets have been labeled as ‘O’, the unlabeled ones are NO).

**Project #12, Supporting/Opposing Proud Boys**

**Supporting Tweets (Y)**

*@sav\_says\_ Does the Proud Boys have a Go Fund page? I canâ€™t be there but Iâ€™d love to support them!*

**Opposing Tweets (N)**

*Seems like proud boys are getting wrecked today in Michigan*

**Project #13, Detecting Offensive Content**

Sarcasm and offensive language is used by both the supporters and detractors of Proud Boys.

**Offensive Tweets**

*F\*ck the Proud Boys. Seriously. They're horrible thugs and it's abhorrent that nothing happens to them when they act out but everything happens to those who just want to be able to live their lives without worry.* [*https://t.co/m7BeE7ROeu*](https://t.co/m7BeE7ROeu)

*@MrAndyNgo So happy Proud Boys kicked their asses.*

**Non Offensive Tweets**

*@MrAndyNgo Go proud Boys!*

*@tedwheeler @JoAnnPDX hereâ€™s the PPB allowing the Proud Boys to skirt through Portland, while obstructing journalists from doing their job. And among the Proud Boys, multiple-charged assailant, Tusitala â€œTinyâ€ Toese. Heâ€™s violated every probation.*

**References**

1. <https://en.wikipedia.org/wiki/Proud_Boys>
2. <https://www.washingtonpost.com/nation/2020/08/22/portland-police-far-right-protest/>
3. <https://www.cnn.com/2020/08/15/us/far-right-rally-proud-boys-michigan/index.html>
4. <https://www.nytimes.com/2019/10/29/reader-center/proud-boys-antifa-fight-video.html>
5. <https://abcnews.go.com/US/facebook-takes-proud-boys-american-guard-accounts-connected/story?id=71286604>
6. <https://www.cnn.com/2020/08/14/media/qanon-news-coverage/index.html>
7. <https://www.rt.com/usa/498753-proud-boys-antifa-blm-portland/>