Identifying and Antisemitism and Hate Speech on Twitter

Jacob Levine

**Capstone Problem and Rationale**

*Introduction to Data Mining*

Data Mining is a rapidly advancing area of Computer and Data Science that involves finding anomalies and patterns within large datasets in order to predict outcomes. Although this sounds relatively straightforward, the field is a relatively complex one that is constantly growing in its scope and its application and depth of research. That said, data mining projects can have big impacts on the ways we function within society and behave in day-to-day life. When choosing another Capstone, I wanted to select something that was not only salient to current happenings in the world but was also something I am personally interested in. One of the most controversial areas of data collection and digital ethics is in social media and internet platforms, specifically due to the current political climate and ethical underpinnings of big data collection and moderation by social media companies. Specifically, I wanted to use Twitter as the source of my project because of its immense amount of user-generated data in accessible and concise formats. Twitter is one of the most popular social media sites and attracts an incredible amount of traffic on a daily basis. In fact, as of Q1 2019 Twitter had more than 330 million monthly active users.[[1]](#footnote-1) Since there is such a huge amount of data being collected in tweets every day, I immediately recognized the Twitter platform as one that could benefit from analysis through Data Mining algorithms and analysis.

*Recent Events*

Recently Big Tech has been in the hotseat with Congress and in the media, as social media and tech moguls attempt to defend themselves against questions and investigations into the legality and ethical standards of many practices these huge tech firms engage in. Twitter’s former CEO Jack Dorsey has spent the past year in and out of Congressional hearings and has stood by his stance that total moderation of the platform isn’t realistically feasible.[[2]](#footnote-2) There are so many tweets going out on a daily basis and companies like Twitter want to prioritize profitability, not wasting resources on moderation when they don’t need to. Currently, Twitter uses fact-checking on some hot-button issues (such as the presidential elections or COVID-19 pandemic) to make users aware of false postings/fake news online. In doing so, however, the company has managed to anger both sides of the political spectrum, as Republican lawmakers call Twitter’s policies discriminatory and are frustrated for the apparent favoring of liberal viewpoints online and overstepping of regulatory measures which they believe completely bias the app. On the other side, Democrats are pushing for more regulation, citing issues like the spread of hate speech and misinformation. It is true that in recent years, Twitter has been a hot bed for the spread and encouragement of hateful content, causing lots of controversy and questions as to the proper and *consistent* application of first amendment rights (and exceptions) on the platform. This is a part of the larger debate surrounding Section 230 internet regulation protections. My project focused on this hateful content and aimed to help study the solutions that could be applied which is vital to understand in the context of my research project as the problem I intend to solve is very closely tied to the Section 230 debate.

In just the past month, Tesla and SpaceX founder Elon Musk bought Twitter for $44 billion, pledging to bring complete freedom of speech to the platform. Although many fear that he will dangerously lower the site’s content moderation policies, it remains to be seen how he will respond to the encouragement of violence online.[[3]](#footnote-3) For the sake of this project, keeping track of changes made to Twitter under new leadership will be increasingly important.

*Hate Speech and Antisemitism on Twitter*

One of my underlying personal motivations for this project was actually that the problem of hate speech, and even more severe threats of violence or implications of danger, are becoming more and more prevalent in shareable digital spaces. I have personally witnessed this type of thing occur and want to make sure that we protect freedom of speech and expression on the internet, while still protecting against crime and predatory violence.

Twitter is unique in that it allows for relatively short snippets of information, opinions or other sharable data to be shared at incredibly high rates to an incredibly large population. The nature of the platform in many ways constitutes the ultimate “idea spreader” and is designed to do so. This has caused many investigations as to the broader effects of trending topics on Twitter, specifically to the implications of hate speech on Twitter to real world hate crimes or violence. In fact, a recent study carried out by researchers at NYU showed a pretty clear correlation between these to phenomena. The researchers analyzed more than half a billion tweets out of a total of 100 cities in the US (from huge cities to small towns) and found that the amount of tweets which contained targeted racism frequently correlated with the number of racially motivated hate crimes reported in those respective cities.[[4]](#footnote-4) This study and others like it went a long way to show the consequences unchecked hateful and/or violent content can have.

Recently, Twitter has brought some change along in order to partially remedy these patterns. For example, in June 2020 Twitter updated its hate speech policy to include links to external hateful or violent content.[[5]](#footnote-5) In December 2020, the company went even further to forbid “language that dehumanizes people on the basis of race, ethnicity and national origin” and “will also continue to surface potentially violative content through proactive detection and automation”.[[6]](#footnote-6) While this may seem helpful, not everyone agrees that these changes are practically relevant or made in good faith. A research study published as recently as August 2021 titled “Gatekeepers of toxicity: Reconceptualizing Twitter’s abuse and hate speech policies” derided the platform for becoming a “breeding ground” for hate speech and permissible of encouragements to violence in the name of free speech.[[7]](#footnote-7) Specifically, it seems that Twitter’s policies and technological architecture both follow a system of “nondeterrent enforcement”.

I was originally made aware of the dangers of hate speech after personally interacting with it early on in my life when I got my first social media and Twitter accounts. Growing up with a Jewish background, I experienced firsthand rises in antisemitism over the recent years and Twitter was, without a doubt, one of the places that this type of discriminatory/aggressive content first took off and became popular. The severity of hateful attacks in our country, specifically those based in antisemitism has quickly been climbing over the years. In fact, one of every four American Jews report that they have directly experience Antisemitism.[[8]](#footnote-8) Furthermore, Jewish people made up nearly 2/3rds of all religious hate crimes in 2019 and 2020 despite being only 2% of the population.[[9]](#footnote-9) Studying antisemitism was particularly important to me because, unlike studies on some major forms of online abuse like racism, cyber-bullying and sexism, online antisemitism present on the web communities has not been studied in much detail from a machine learning perspective. This calls for studying online antisemitism in greater depth so as to protect the users from online/real world hate crimes.[[10]](#footnote-10) Hopefully this project can help to serve as an example for and justification for why further research and dataset creation is needed in regard to fighting antisemtitism and other forms of hate online.

*Data Mining: Text Analysis Concepts*

Text analysis is a machine learning technique that we can use to automatically record and use patterns from text data to produce prediction and classification outputs. Many companies already use text analysis in web scraping or data collection contexts to digest large amounts of data or process a large amount of documents/paperwork, turning these resources into actionable and valuable insights.[[11]](#footnote-11) Text analysis is meant to be used to go over large amounts of text data and pull information using a computer program from the text to then analyze further.

Natural language processing (or NLP) is a machine learning technique used often in text analysis that enables computers to try and simulate a human-like understanding and breakdown of text material. Of course, there are different types of NLP and many different ways to do text analysis, but the most straightforward and most popular is likely text classification, which includes its own family of machine learning strategies including sentiment analysis, topic modeling, language detection and intent detection. Each of these sub models are similarly inclined, focusing on taking a string of text into the program as an input and producing an output that categorizes the text based on the respective traits of sentiment, topic, or intent.

Past studies involving the use of machine learning algorithms to detect hate speech in massive text datasets have identified a variety of challenges that could begin to get in the way of collecting accurate results. These difficulties primarily revolve around differing definitions on what constitutes hate speech and limitations of data availability (or of access to data that can actually be used in the study conveniently). There are many different ways that studies have chosen to account for this in the past, including a focus on a specific working definition or a focus on transparency in the machine learning algorithm’s decisions in order to create a manual appeal process for hate speech identification or censorship.[[12]](#footnote-12) The challenge is to ensure that there is a distinction made between real hate speech and offensive content.

The project I undertook, while not magically solving all of these problems or issues in hate speech research, takes many of these concerns into account in some regard. I’ve also taken feedback and insight from the studies I have read about to carry out my own project in an effective and efficient way so that I can get some meaningful information and analysis that tries to avoid obvious mistakes, which I have been warned of, and produce something relevant to the greater body of work with this topic given limited time constraints and having only *just* learned about Data Mining as a novice in the subject.

**Objectives/Deliverables**

Objectives

The goal of this capstone was to learn about the most effective ways to identify hate speech in vast amounts of twitter data and develop an effective methodology to produce meaningful results from this data. Essentially, the main objective was to begin a study on the machine learning process or set of algorithms that can identify antisemitic tweets and categorize them into meaningful and actionable areas. This project helped to learn about the moderation of threatening content on twitter and provide the ability for us to learn more about antisemitism as it exists on the internet and how it provokes violent action against religious and ethnic Jewish populations. It taught about the types of antisemitism that exist. It helped to determine the best ways to extract and eventually analyze large amounts of Tweet data or other social media content. It also gave important insights as to the effect of online speech on real world happenings.

Deliverables

* Research Project Report
* Research Project Data
* Project Methodology and Code
* Information and takeaways based on results

**Project Description**

Narrative

I began a project to produce a machine learning architecture that inputs Twitter data and utilizes a variety of different types of machine learning algorithms to identify antisemitic tweets and categorize them based on appropriate standards of antisemitism and hate speech. During this semester, I took the first steps in this project by gaining access to a Twitter Developer account, learning how to use the Twitter API, and beginning the curation of datasets for use in deep learning algorithms. I will optimized the machine learning architecture based on comparing the most effective ways to generate meaningful results and discussed takeaways with cross-disciplinary experts.

Timeline

This was an incredibly ambitious project that faced many unforeseen developments and obstacles. Regardless, I was able to make progress and set a foundation for the future of this project:

1. Sprint 1:  1/24-2/7

* Developed an architecture plan
* Reached out to relevant experts in other fields

1. Sprint 2: 2/7-2/21

* Background Research
* Application to the Twitter API

1. Sprint 3: 2/21- 3/7

* Gained access to the Twitter API
* Experimented with tools to access the Twitter API
* Developed dataset keyword list to extract meaningful tweets

1. Sprint 4: 3/7- 3/21

* Used dataset keyword list to begin extraction of Tweets for the dataset
* Expanded and focused Postman extraction of Twitter API

1. Sprint 5: 3/21- 4/4
   * + - Finished dataset extraction from Twitter API
       - Worked on familiarizing myself with relevant Python packages
2. Sprint 6: 4/4-4/18
   * + - Downloaded appropriate python packages for dataset curation
       - Merged datasets to single dataset
       - Started text preprocessing (cleaning, lemmatization, tokenization etc)
3. Sprint 7 4/18-4/29
   * + - Finished text preprocessing (cleaning, lemmatization, tokenization etc)
       - Started feature extraction and embedding
       - Prepare final project report presentation and paper

Resources

For this project, I used a wide variety of resources and tools. I used the Twitter API which I accessed through Postman. I used HTML at certain occurrences to do this. I also used Python including packages such as NumPy, Pandas, Seaborn, Sklearn, NTLK and others. I used Google Colab and PyCharm to write code and analyze the dataset. I also prepared the use of certain tools for later stages of the project, such as Word2Vec and genism models.

**Conclusion**

Over the course of this semester, I’ve learned a lot about the overall architecture needed to accomplish the end goals of this project. Within the GitHub I’ve created, are multiple research papers that I used in my background research. Additionally, I included the datasets in both .json and .csv forms, along with the code that I compiled to clean and preprocess the datasets in preparation for clustering methods.

The dataset was extracted with keywords ranging across a variety of antisemitic tropes. The various tweets were then merged together using online tools for file processing and plugged in to Python code in a series of packages to clean it from special characters, stop words and then split by word into a corpus for further analysis.

Future work should entail the expansion of the Twitter dataset and application for a higher level of authorization on the Twitter API. Subsequently, future research would work on applying clustering and running data visualization to analyze the relationships present within the Twitter data. Variations can be made to clustering algorithms, visualizations and datasets themselves to continue to investigate these relationships.

This project, in the long term, could contribute significantly to the discourse surrounding internet regulation and could help give researchers and activists a framework for combatting antisemitism online.

1. Molina, Brett. “Twiter overcounted active users since 2014, shares surge on profit hopes”, *USA Today*, 26 Oct 2017, Retrieved from https://www.usatoday.com/story/tech/news/2017/10/26/twitter-overcounted-active-users-since-2014-shares-surge/801968001/ [↑](#footnote-ref-1)
2. The New York Times, “Zuckerberg and Dorsey Face Harsh Questioning From Lawmakers”, *The New York Times*, 17 Nov 2020, Retrieved from https://www.nytimes.com/live/2020/11/17/technology/twitter-facebook-hearings [↑](#footnote-ref-2)
3. The Wall Street Journal, “Elon Musk Buys Twitter: The Big Picture”, 02 May 2022, Retrieved from https://www.wsj.com/articles/elon-musk-buys-twitter-the-big-picture-11651530974 [↑](#footnote-ref-3)
4. Leskin, Paige. “A new study found a link between the number of racist tweets and real-life hate crimes in 100 US cities”, Insider, 26 Jun 2019, Retrieved from https://www.businessinsider.com/twitter-racism-hate-speech-linked-real-life-hate-crimes-study-2019-6 [↑](#footnote-ref-4)
5. Hutechinson, Andrew. “Twitter Updates Hate Speech Policy to Include Links to ‘Hateful’ Content”, *Social Media* Today, 28 Jul 2020, Retrieved from https://www.socialmediatoday.com/news/twitter-updates-hate-speech-policy-to-include-links-to-hateful-content/582482/ [↑](#footnote-ref-5)
6. Al Jazeera News, “Twitter expands hate speech rules to include race, ethnicity”, 3 Dec 2020, Retrieved from https://www.aljazeera.com/news/2020/12/3/twitter-expands-hate-speech-rules-to-include-race-ethnicity [↑](#footnote-ref-6)
7. Konikoff, Daniel. “Gatekeepers of toxicity: Reconceptualizing Twitter’s abuse and hate speech policies”, *Wiley Online* Library, 29 Aug 2021, Retrieved from https://onlinelibrary.wiley.com/doi/10.1002/poi3.265 [↑](#footnote-ref-7)
8. Hernandez, Joe. “1 in 4 American Jews say they experienced antisemitism in the last year”, *NPR*, 26 Oct 2021, retrieved from https://www.npr.org/2021/10/26/1049288223/1-in-4-american-jews-say-they-experienced-antisemitism-in-the-last-year [↑](#footnote-ref-8)
9. Sharon, Jeremy. “Antisemitism: Jews target of 58% of all religiously motivated hate crimes in US”, *The Jerusalem Post*, , 31 Aug 2021, Retrieved from https://www.jpost.com/diaspora/antisemitism/antisemitism-jews-target-of-58-percent-of-all-religiously-motivated-hate-crimes-in-us-678228 [↑](#footnote-ref-9)
10. Chandra et al. “’Subverting the Jewtocracy’: Online Antisemitism Detection Using Multimodal Deep Learning”, *Conference 17 July 2017 Washington. DC.,* 18 Jun 2021, Retrieved from https://arxiv.org/pdf/2104.05947.pdf [↑](#footnote-ref-10)
11. Monkey Learn, “Text Analysis”, Accessed 14 Dec 2021, Retrieved from https://monkeylearn.com/text-analysis/ [↑](#footnote-ref-11)
12. MacAvaney et al. “Hate speech detection: Challenges and solutions”, *Plos* One, 20 Aug 2019, Retrieved from https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0221152 [↑](#footnote-ref-12)