Machine Learned Text Analysis on a Dataset of Censored Tweets

William Graif, Ethan Waldman

Abstract

Abstract.

1 Introduction

As traditional media has been digitized and moved online, governments around the world have implemented new strategies to censor unwanted content in their region. Many of these approaches involve a government filtering content by ip address or by analyzing traffic, but social media platforms have given rise to newer methods of content moderation. Governments can now send requests directly to social media platforms requesting for certain content to be blocked for their citizens. While the takedown process itself is known, very little is known about the criteria countries use for taking down content on online platforms. We decided to focus our project on the social media platform, Twitter. We found a publicly accesisble dataset of 583,000 tweets [2] that were censored in various countries between 2012 and 2020. In this paper, we use these tweet IDs to download the full tweet data from Twitter's api for all 583,000 tweets. This data contains the full text of the tweet as well as metadata including the countries that tweet is censored in. Using this information, we analyze the text using several machine learned models for named entity recognition and sentiment analysis. We then analyze a random sample of twitter data to compare with our censored tweets. Ultimately we hope that the differences found in this comparison will illustrate the criteria certain countries use to determine which tweets to censor.

2 Data Collection

We used the python library Twarc in order to download tweets from twitter's api. The dataset [2] of Tweet IDs is a single csv file with one Tweet ID per line. Twarc's 'hydrate' command takes in a list of Tweet IDs and returns a json file containing all of the downloaded tweets, so we used this with the published dataset. Once we obtained this tweet data, we converted the json into a dataframe using the python library pandas. We discovered that while several countries were censoring tweets, many only had one or two blocked tweets in the dataset. We decided to limit our project to include tweets censored in France, Germany, India, Russia, and Turkey, seeing as they all had at least several thousand samples. We also found that a significant portion of the data consisted of tweets that were not written in English. This posed an issue for our machine learned models, which were trained only on English datasets. We decided to remove the non-English tweets to account for this. Limiting our data to only include English tweets that were censored in these five countries brought our total tweet count down from approximately 583,000 to 49,000.

In order to collect a random sample of Twitter data to compare to our censored tweets, we again used Twarc. First, we grouped the tweets by the country in which they were censored. Then, we counted the number of censored tweets per country, and attempted to request a random sample of the same number of tweets from 2012 to 2020 from Twitter. Unfortunately, Twitter's built-in historical sampling feature is restricted by API access level, and we were thus unable to use it. To circumvent this, we then attempted to generate random timestamps between 2012 and 2020 and search for tweets published at those times. Unfortunately, we were again limited by API access level, rendering us unable to search for tweets based upon these randomly generated timestamps. We decided to use Twarc's 'sample' command to collect our tweets. This command listens to Twitter's streaming endpoint and gets a random sample of tweets being published in real time. This enabled us to gather a random sample of tweets to match each country's respective censored tweet count in our dataset, but since it used Twitter's streaming endpoint, all of the tweets were recently published, so they did not fall in the same time period as our censored tweets. This is potentially a confounding variable when we compare the censored tweets to the uncensored tweets for a given country: we would like to attribute any major differences in the sentiment and topics between these two categories to the country's censorship, but they may also be attributable to different topics and sentiments that people are truly expressing at various points in time. We ran this collection process on March 15th at 12:00 AM CST, so all of our random samples were tweets published around that time.

3 Analysis

For our named entity recognition analysis, we opted to use Flair's large 4 class model, which is based on FLERT [3]. For sentiment analysis, we used a roBERTa-base model trained on 58 million tweets, based on the TweetEval benchmark [1] for tweet classification. Before analyzing our tweets, we had to preprocess the text to remove usernames, urls, and other unecessary information. For our preprocessing, we used the library tweet-preprocessor. We iterated through all of the data we collected and used the library's clean() function on the text for every tweet object. After this process, we iterated through all of the data again, this time performing our analysis using the models described previously. Flair's NER model was able to detect organizations, locations, and people mentioned in tweets. It also picked out some other words that did not fit those criteria, but were still determined to be important and labeled them as miscellaneous. An example of a detectable location is 'Ukraine', while the word 'Ukrainian' was labelled as miscellaneous. We saved the top 10 most commonly found people, places, organizations, and miscellaneous items for the censored tweets as well as the sample data for each country. Our sentiment analysis returned three numbers between zero and one for every tweet. Each one was a rating for the positivity, neutrality, or negativity of the text. For example, a strongly negative tweet would have positivity and neutrality scores close to zero, while the negativity score would be close to 1. We saved these three scores for every tweet, and then averaged out the scores for the censored and random data for each country. We also performed a t-test on the positive, negative, and neutral scores of the random and censored data to compare them and determine if there is a statistically significant difference between the two datasets.

4 Results

First, here is a table showing the average values for our sentiment analysis scores as well as the p-values calculated from our t-test. The column Censored shows the average score for our censored data while the column Sample shows the average score of our random data. The P-Value column shows the P-Value calculated from a t-test comparing the censored and sampled data for a given score category. After this table, there is another that displays the top ten items from all of the classes of our named entity recognition for both of our datasets. We begin with the information for France:

Score Category	Censored	Sample	P-Value
Positive	0.026	0.0404	2.2e-13
Neutral	0.211	3.52	0.768
Negative	0.763	0.75	0.089

Censored Topics:

People	Locations	Organizations	Misc
'Trump'	'America'	'CAIR'	'Jews'
'Allah'	'Germany'	'ISIS'	'Islamic'
'God'	'France'	'HAMAS'	'Muslim'
'Merkel'	'Europe'	'Muslim Brotherhood'	'Muslims'
'Salvini'	'UK'	'Twitter'	'Islam'
'Ilhan Omar'	'Israel'	'FBI'	'Christians'
'Muhammad'	'USA'	'Hamas'	'Sharia'
'Hitler'	'U.S.'	'EU'	'Jewish'
'Obama'	'Italy'	'CNN'	'German'
'Rashida Tlaib'	'US'	'Libs'	'Christian'

Uncensored Topics:

People	Locations	Organizations	Misc
'Ruto'	'India'	'Vogue Koreas'	'Russian'
'jungkook'	'USA'	'INOX'	'Indian'
'God'	'Russia'	'RT'	'Pico'
'Tae'	'RUSSIA'	'NFT'	'Group of the year'
'Putin'	'New York'	'Discord'	'Fandom of the year'
'Tom Brady'	'Europe'	'Girl'	'return of superman'
'Jisoo'	'Central America'	'YTIn'	'Barbies'
'chan'	'Philippines'	'Youku'	'Twitter'
'ekneuzmo'	'US'	'skz'	'Breaking Dawn'
'sheytan'	'Indy'	'Mel0n'	'Dusty Pink'

Germany:

Score Category	Censored	Sample	P-Value
Positive	0.029	0.038	1.88e-22
Neutral	0.219	0.842	1.14e-5
Negative	0.753	0.756	0.265

Censored Topics:

People	Locations	Organizations	Misc
'Trump'	'America'	'CAIR'	'Islamic'
'Allah'	'Germany'	'ISIS'	'Islam'
'God'	'France'	'Muslim Brotherhood'	'Jews'
'Merkel'	'Europe'	'HAMAS'	'Muslim'
'Mohammad'	'UK'	'Twitter'	'Sharia'
'Hitler'	'Israel'	'FBI'	'Muslims'
'Salvini'	'USA'	'EU'	'Christians'
'Muhammad'	'U.S.'	'Hamas'	'German'
'Obama'	'Italy'	'UN'	'Jewish'
'Ilhan Omar'	'US'	'CNN'	'Christian'

Uncensored Topics:

People	Locations	Organizations	Misc
’God’	’Russia’	’RT’	’IDR’
’KAMLE JODI’	’Ukraine’	’JT’	’Group of the year’
’Putin’	’China’	’US’	’Russian’
’JT’	’UK’	’India’	’American’
’Trump’	’India’	’China’	’Ukrainian’
’LISA CHIVAS’	’US’	’Kyiv’	’Fandom of the year’
’Tom Brady’	’MINS’	’Kashmir’	’White Day’
’Louis’	’Turkey’	’USA’	’Indian’
’Biden’	’Texas’	’America’	’ANGELIC’
’jungkook’	’USA’	’Uganda’	’Kashmir Files’

India:

Score Category	Censored	Sample	P-Value
Positive	0.031	0.042	4.77e-6
Neutral	0.214	2.76	0.264
Negative	0.755	0.752	0.741

Censored Topics:

People	Locations	Organizations	Misc
’Modi’	’India’	’RSS’	’Indian’
’Trump’	’Pakistan’	’UN’	’Kashmiri’
’God’	’Kashmir’	’RAW’	’Muslims’
’Bill Gates’	’US’	’IK’	’Kashmiris’
’Allah’	’UK’	’Twitter’	’Muslim’
’IK’	’Israel’	’CFL’	’Indians’
’Pak’	’USA’	’CIA’	’Pakistani’
’Zardari’	’U.S.’	’AJK’	’Zionists’
’Gates’	’Italy’	’WHO’	’Israeli’
’Dr MSG’	’US’	’TTP’	’Afghan’

Uncensored Topics:

People	Locations	Organizations	Misc
’Jesus’	’Europe’	’Vogue Korea’	’Russian’
’Garibdas Ji’	’Kyiv’	’JISOO’	’IDR’
’Louis Tomlinson’	’INSANE’	’Vogue Koreas’	’Sorry’
’Brit’	’DALAMPASIGAN’	’YouTube Theater’	’RELEASENCT DREAM - The nd Album’
’God’	’Melitopol’	’Congress’	’Glitch Mode’
’JJ Guirella’	’Miami’	’FBA’	’Spider-Man: No Way Home Toms Journey’
’Justin Bieber’	’Ever’	’Facebook’	’BE WITH ME’
’Elon Musk’	’london’	’RT’	’Scandinavian’
’Tom’	’Ghana’	’Nice’	’Viking’
’John Cleese’	’MIAMI’	’BUMZU’	’Pro-Kremlin’

Russia:

Score Category	Censored	Sample	P-Value
Positive	0.029	0.038	2.23e-20
Neutral	0.213	0.319	0.003
Negative	0.758	0.759	0.851

Censored Topics:

People	Locations	Organizations	Misc
’Allah’	’US’	’Pinnacle’	’Muslims’
’Trump’	’U.S.’	’Liverpool’	’Muslim’
’Putin’	’Russia’	’Arsenal’	’Islam’
’Neymar’	’Pakistan’	’Real Madrid’	’Russian’
’Bismillah Al-Rahman Al-Raheem’	’England’	’Chelsea’	’Chechen’
’Jurgen Klopp’	’Europe’	’Bovada’	’Dragons Slot’
’Khilafah’	’India’	’Barcelona’	’Islamic’
’Naveed Butt’	’Chechenia’	’Manchester United’	’Premier League’
’Arsene Wenger’	’Ukraine’	’Hizb ut Tahrir’	’Chechens’
’Pep Guardiola’	’UK’	’Microgaming’	’Texas Holdem’

Uncensored Topics:

People	Locations	Organizations	Misc
’God’	’Ukraine’	’RT’	’IDR’
’Putin’	’Russia’	’JT’	’Russian’
’JISOO’	’China’	’HRS’	’Ukrainian’
’KAMLE JODI’	’UK’	’NFT’	’Group of the year’
’JT’	’India’	’Congress’	’Fandom of the year’
’Sakura’	’US’	’HOURSRT’	’American’
’Louis’	’MINS’	’BTS’	’Super Tuna’
’Jin’	’Turkey’	’The Group’	’White Day’
’Allah’	’Texas’	’Twitter’	’Turning Red’
’Tom Brady’	’USA’	’IG’	’YouTube’

Turkey:

Score Category	Censored	Sample	P-Value
Positive	0.026	0.039	1.32e-219
Neutral	0.224	0.302	1.24e-31
Negative	0.749	0.752	0.1002

Censored Topics:

People	Locations	Organizations	Misc
’Erdogan’	’Turkey’	’ISIS’	’Turkish’
’Trump’	’Syria’	’SDF’	’Kurdish’
’Erdogans’	’Afrin’	’YPG’	’Kurds’
’Erdoan’	’US’	’daesh’	’Turk’
’Assad’	’Efrn’	’PKK’	’Syrian’
’Gulen’	’Raqqa’	’YPJ’	’Turks’
’Fethullah Gulen’	’Rojava’	’Daesh’	’Ezidi’
’Putin’	’Kurdistan’	’HDP’	’Turkeys’
’Anna Campbell’	’UK’	’EU’	’Russian’
’Enes Kanter’	’Manbij’	’AKP’	’German’

Uncensored Topics:

People	Locations	Organizations	Misc
’Putin’	’Ukraine’	’RT’	’Russian’
’God’	’Russia’	’JT’	’IDR’
’JT’	’US’	’HOURSRT’	’Ukrainian’
’Louis’	’India’	’BJP’	’Group of the year’
’Tom Brady’	’China’	’Twitter’	’Fandom of the year’
’JISOO’	’Kyiv’	’Congress’	’Indian’
’Trump’	’USA’	’BTS’	’American’
’Lisa’	’Kashmir’	’HRS’	’Hindu’
’jungkook’	’UK’	’NFT’	’Turning Red’
’Biden’	’SEOUL’	’IG’	’White Day’

5 Related Work

[1] [2] [3]

6 Conclusion

References

[1] Francesco Barbieri, José Camacho-Collados, Leonardo Neves, and Luis Espinosa Anke. “TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification”. In: *CoRR* abs/2010.12421 (2020). arXiv: `2010.12421`. URL: `https://arxiv.org/abs/2010.12421`.

[2] Tugrulcan Elmas, Rebekah Overdorf, and Karl Aberer. “A Dataset of State-Censored Tweets”. In: *CoRR* abs/2101.05919 (2021). arXiv: `2101.05919`. URL: `https://arxiv.org/abs/2101.05919`.

[3] Stefan Schweter and Alan Akbik. *FLERT: Document-Level Features for Named Entity Recognition.* 2020. arXiv: `2011.06993 [cs.CL]`.