Project Report

USE OF AI TO PREDICT MALICIOUS CONTROVERSIES WITHIN POPULATION

Sanket Ghanmare
Foundations of Artificial Intelligence, Fall-2020

Contents

Problem my AI model will try to solve	Page 3
Motivation	Page 3
Examples of malicious controversies	Page 4 & 5
Problem statement & High-level Project Overview	Page 5
Phase 1, Phase 2 descriptions along with diagrams	Page 7, 8 & 9
Dataset using Twitter API	Page 10
Algorithm, Model Design, Step 1, 2 and 3 of Algorithm	Page 11, 12 & 13
Activity diagram of Malicious Controversy Detector	Page 14
Result given by phase 1 model and description of Result with details.	Page 15, 16 & 17
Bag of Word Model and steps I use to implement it.	Page 18 & 19
Experimental setup details & hypothesis for Phase 1 and Phase 2	Page 20 & 21
What worked, what didn't, Future direction, Changes I did, etc.	Page 22 & 23
Advice for CS4120 future students.	Page 24
References & Articles	Page 25& 26

Use of AI to predict malicious Controversies within Population

Problem my AI model will try to solve: -

My theoretical model will try to address this issue of controversies which can have a serious negative impact and provide a possible solution to detect or predict such malicious controversies and alert authorities in advance to take necessary steps to stop any unfortunate event at a particular Geo Location by tracing the source of perpetrators who started it and the regions where it can have a serious impact.

Motivation: -

My motivation for this project comes from my deep desire to use AI to predict controversies among people and predict (or detect) whether any controversy has a negative impact which may give birth to dangerous events like Riots, loot, War, Terrorism, or other malicious activities. In the last couple of decades, the internet has evolved a lot and along with the internet many other technologies.

Controversial issues may be questions, subjects, or problems that can create a difference of opinion. They can include issues that may have political, social, environmental, or personal impacts on pupils and or the wider community: locally, nationally, or internationally. Often, they have no easy answer, in part, because solutions may be based on an individual's values and beliefs.

In today's world, anyone who has access to the internet can create their private account on the vast majority of Social Networking sites which help that account used to connect and contact people across the globe. It has both Pros and Cons. If someone or a group of people decides to use this to harm and destroy the integrity of a country or a particular community then it is very much possible to do so. Not only social media but many elements in society can create controversies in the form of fake news, edited images, etc.

Below are some famous examples of malicious controversies:

During a recent campaign in Syria, the Turkish Armed Forces borrowed a page from ISIS's playbook, using social media to instill a sense of unopposable force.

Example: - 1

The Islamic State's careful audiovisual engineering hints at a future of war propaganda that will lean almost entirely on evocative and shareable images—everything from doctored photographs to video screenshots to infographics. ISIS militants have discovered, as marketing experts have long known, that compelling imagery matters far more than any accompanying text in determining whether or not something goes viral. Indeed, when the Turkish military launched an August offensive into Syria to sweep ISIS militants from its border, it cribbed many of the very same online tactics, creating a Twitter account for the operation that pushed out everything from soldier selfies to dramatic, staged videos of commando raids.(Source: Atlantic)

The Israel Defense Forces are very active on social media. During conflicts, "war rooms" of Israeli soldiers and students vie with Palestinians to shape global perceptions.

Example: - 2

These questions are no longer so fanciful. In recent conflicts, Israel has established Hasbara (the Hebrew word for "explanation") war rooms, filled with university students and soldiers who tangle with Hamas and Palestinian sympathizers over what, exactly, is going on in their perpetual war. The scale of this online jousting is astounding. During the 2014 flare-up in Gaza, for example, the two sides' competing hashtags, #GazaUnderAttack and #IsraelUnderFire, racked up some 5 million uses. The Wikipedia page about the conflict has been edited and reedited more than 7,000 times. (Source: Atlantic)

Another famous example of malicious controversy which happened very recently:

"Russian social-media users and allied accounts seek to manipulate opinion and sow dissension in enemy states. (Source: Atlantic)."

Above are some of a few examples but there are many such controversial topics such as Gun Control, Abortion, Human rights, etc.

Problem statement which I tried to solve: -

Can I build a theoretical model to detect malicious controversies which can output the Geolocation of perpetrators as well as possible locations where an uneasy situation may take place based on? And predict when that situation may occur based on the analysis of previous criminal records and the motivation behind those crimes in that particular region.

The High-level idea for of Project:

I want to build a model that can classify maliciously controversial topics from current trending topics that are popular all around the world.

They can be present on various platforms such as Social networking platforms (ex: - Twitter, Facebook, etc.), news channels, News Articles, Video streaming sides (ex: - YouTube, etc.)

My model will try to separate these topics and locate the users which posted comments, spread false rumors, etc. on these topics to spread unrest, hate, violence, etc. in the society.

Once my model filters out such controversial topics along with all the perpetrators behind these controversies it will try to locate their Geolocation and based on this Geolocation my model will perform an analysis of all the crimes that previously happened on that Geolocation along with the motivation behind those crimes.

If the motivation behind those crimes is similar or exactly matches with the controversial topic and all the comments, or other content related to the topic then it will alert authority and give them all the necessary data and its overall analysis results.

Note: -

Initially, I planned to work on a theoretical project but with time I decided to divide the entire project into two phases, and I am able to build a model for phase 1 which shows good results for demo purposes. Phase-2 is still theoretical.

Phase 1: - Sort out Malicious Controversial topics from controversial ones.

To complete Phase 1, I studied topics related to Controversies and their effects in conversation.

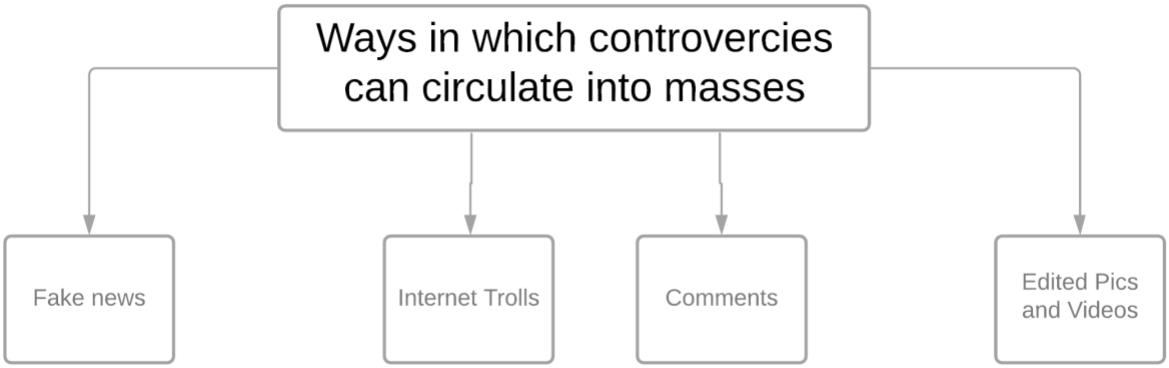


Fig 1: - Different ways for creating and circulating controversies

These are the different ways through which controversies can be created and circulated into the masses.

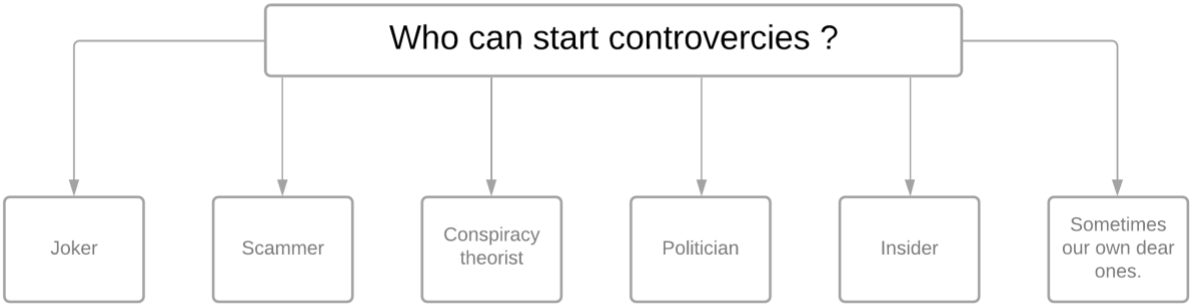


Fig 2: - Different elements who can start controversies.

Phase 2: - ***Predict the location, time, and controversies which has the potential to cause serious damage.***

For Phase 1 I have to build a model to list out controversial topics and plot them using a scatter plot. Each time my model runs it creates the CSV file having columns as follows: Topic, Controversy Rate (CR), Remark, Tweet, and User.

For Phase 2, I performed a theoretical analysis based on the results obtained from Phase 1.

I wanted to build a model that can briefly explain and show how the actual model I tried to solve in this project so for demonstration purposes I used *Twitter API* by creating a student account on the twitter developer website.

Analysis of previous crime records along with time, location, nature of the crime, and the main motivation behind the crime is very important.

Based on the output acquired from Phase 1 which consists of Topic, CR, Remark and Tweets, and User list Object. Phase 2 analysis can start.

In the analysis, the location of Twitter users along with their tweets and Topic of tweets which is remarked as controversial is matched with the previous location of crimes, the main motivation behind those crimes (ex: political, personal gain, substance abuse, looting, etc.), types of crimes (ex: violent crime, hate crime, organized crime, sex crime, cyber-crime, etc.) and the time when that crime took place.

Using the above as features for forecasting the next location where a crime may take place because of a controversial topic. This can be done using a machine learning model along with deep learning techniques.

One such good starting point I found out to do prediction is by studying the way weather forecasting models work. These forecasting algorithms use machine learning techniques i.e., *Bayesian Network and Neural Network*.

Examples of good forecasting algorithm which can be used to implement Phase 2 are Autoregressive Integrated Moving Average (ARIMA), Prophet library which implements additive time series forecasting, etc.

Another good candidate is *the COMPAS algorithm (source: Wikipedia)* which is designed to predict and decide whether the crime will take place or not and also it has much more uses apart from this one use I mentioned although it is not yet perfect but can show promising results if more work is spent on it to improve.

In my project Phase, 2 has a unique set of challenges such as predicting the maliciously controversial topics, the location where this controversy has the potential to do some lethal damage as well as what will be the period these topics will take to reach the masses.

Continuous real-time data as well as monitoring of all the platforms such as social media platforms, news channels, streaming websites, etc. is needed for this model to be successful, or else it won't show a promising result.

Dataset using Twitter API: -

To build the first phase of the model I used the ***Twitter API*** to request the data listed in the list of trending topics. I used *the* ***tweepy library*** to make Twitter API calls.

I wanted the real-time data which is one of the most important key features for the entire project based on which the model will do its analysis and shortlist topics that are highly controversial and needs to be either flagged on online platforms or in some cases need to be removed.

However, since it's a student account Twitter does not allow me to fetch more than 200 tweets per request for every topic which is currently trending and needs to be examined meticulously to avoid unfortunate events in near future either on social platforms or in the real world.

```
tweets = api.get_tweets(query=topic, count=5000000)
```

Fig 3: - Twitter Api Request

Below is the list of trending topics on which I performed Phase 1 of my project.

```
# list of trending topics.
listOfTopics = ['hatred', 'racism', 'war', 'Election', 'Trump', 'cdc vote on vaccine distribution',
                'who gets the vaccine first', 'joe biden', 'terrorism', 'who are considered health care workers',
                '14-day quarantine states', 'soccer', 'corona virus', 'lockdown', 'play station 5',
                'Cinnamon', '#Goya']
```

Fig 4: - Topics list

Algorithm and Model Design: -

Following are the Algorithm that I devised after completing my studies mentioned in Milestone Report 1 of Project.

Step 1 of the Algorithm: -

Build a vocabulary (list of words) of all the words in the training data set.

Match tweet(text) content against vocabulary — word-by-word.

Build a word feature vector.

Plug feature vector into the Naive Bayes Classifier.

Step 2 of the Algorithm: -

Make a list of all the negative comments/ Topics classified by the sentiment analysis algorithm.

Now arrange them according to the number of:

TAC = like + dislikes + comments + emojis(reactions)

TAC = Total Attention Count

Total Attention Count represents how many people around the world paid attention to a comment(discussion) in the form of like, dislike, comment, or emojis, which is classified as negative by sentiment analysis.

Calculate the controversy rate for each topic that has high TAC.

$$CR = \frac{\text{Total Dislikes} + \text{Total Negative Comments} + \text{Total Negative Reactions}}{\text{Total Dislikes} + \text{Total Negative Comments} + \text{Total Negative Reactions} + \text{Total Likes} + \text{Total Positve Comments} + \text{Total Positive Reactions}}$$

Create a list of topics along with their CR value.

Step 3 of the Algorithm: -

Filter out all the topics which have a CR >= 40% in a separate list.

Mark these topics as controversial.

Locate the user Id who posted this topic as well as those user ids who commented on the topic.

Trace the geographical location of these users.

Analyze the previous data related to riots, peaceful protests, loot, shootings or any other type of crimes on these locations

Filter out the most common reasons behind these crimes and create a list of these.

Match the keywords of the Topic and its related comments, videos, news, etc. with the words present in the list and detect how similar are they.

If the matching rate is more than 70% then categorize this topic as malicious and alert authority present at or nearby GeoLoaction to take necessary steps to maintain peace and harmony and provide them entire results (ex: usernames, there comments, location, topic, etc.) acquired after performing in depth analysis.

Note: -

TAC = Total Attention Count

CR = Controversy Rate

*(Inspiration and Source: "**Controversy trend detection in social media**" see reference)*

To design the Model, I used the above algorithm combined with all the steps mentioned in the Dataset section above.

Below is the High level 3 tier architecture of Algorithm: -

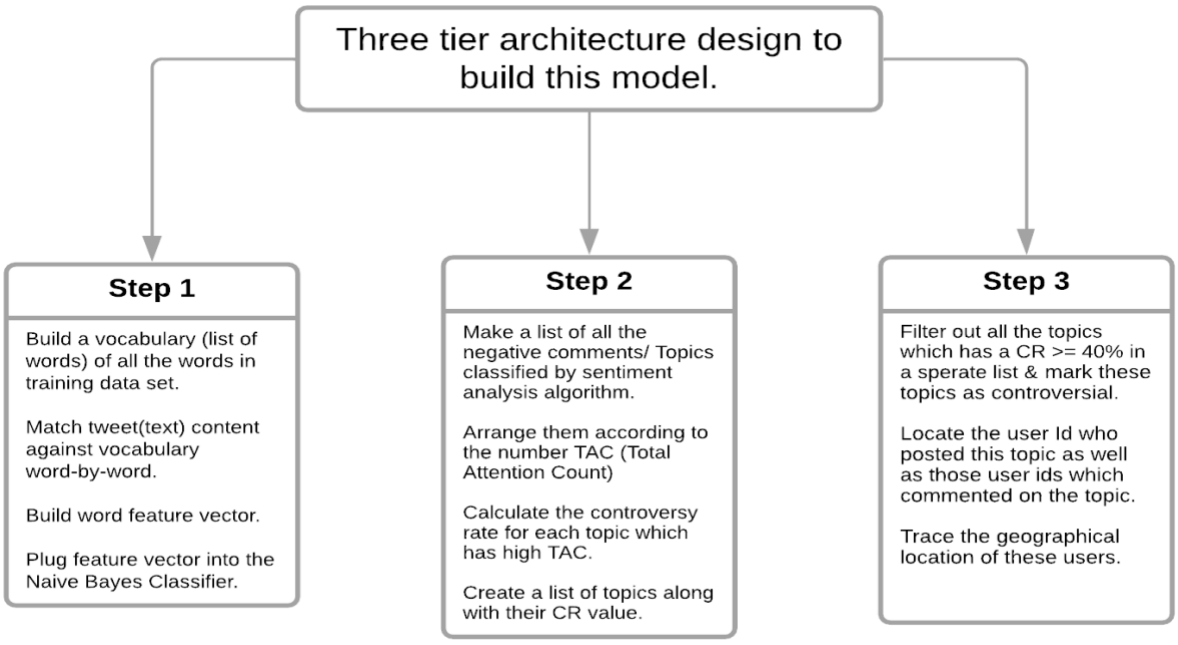


Fig 5: Three tier architecture

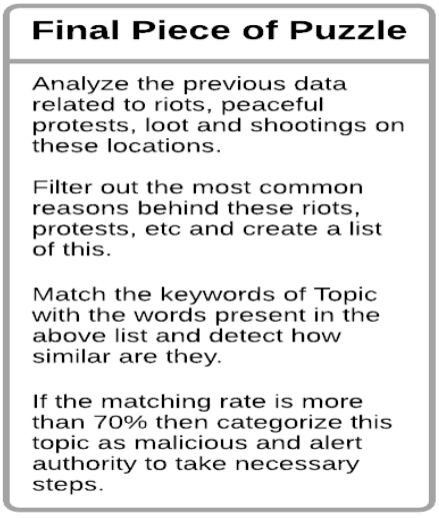


Fig 6: Final step of the model

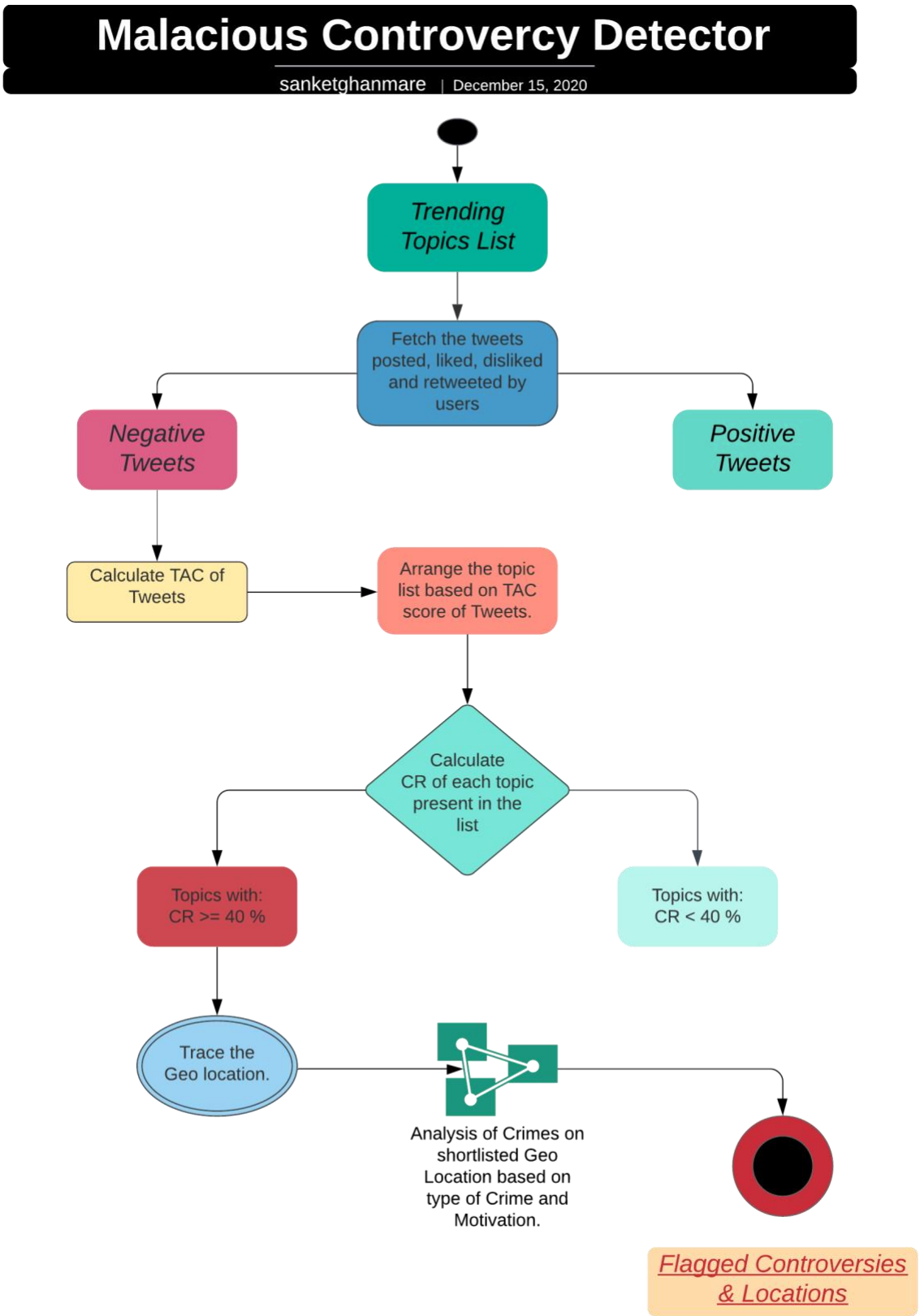
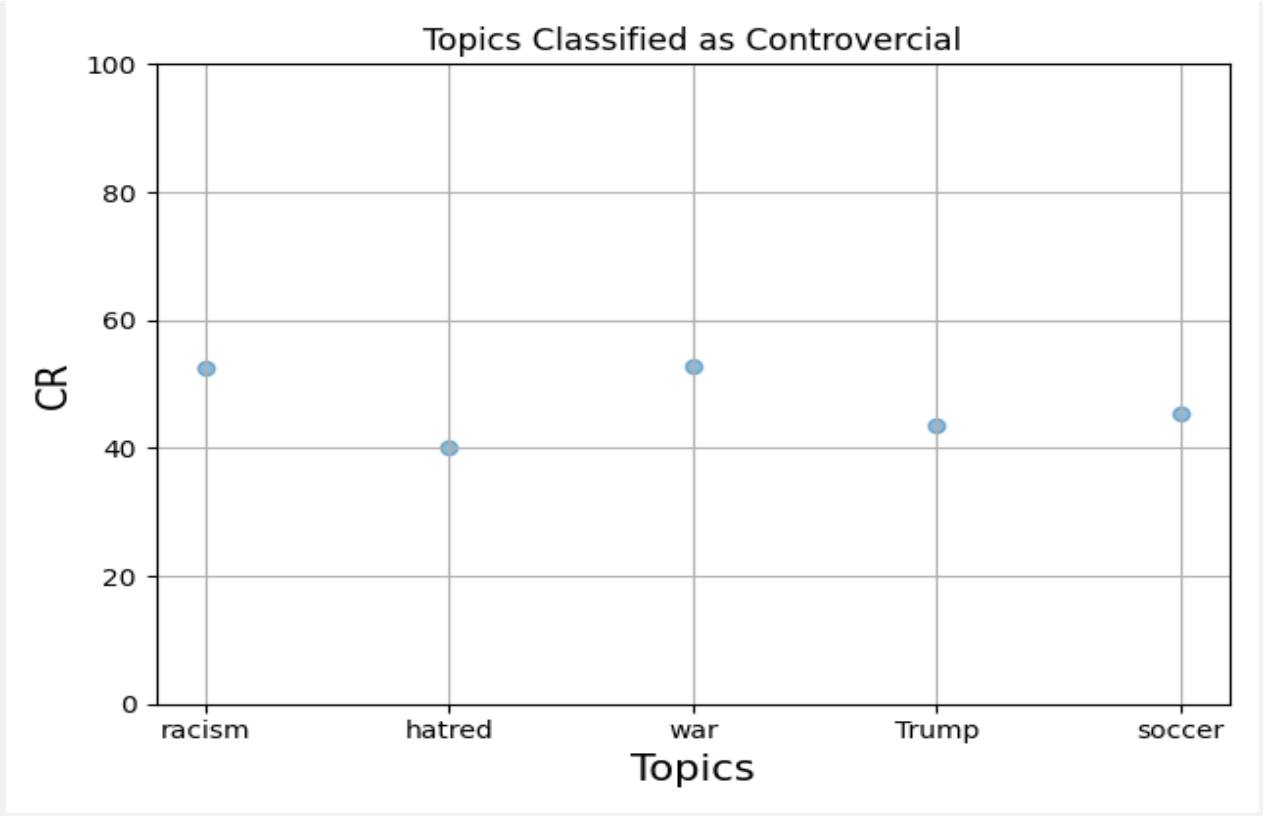


Fig 7: Flow diagram of Algorithm

Following are the various different output given by Phase 1 Python model: -



Plot 1

A	B	C	D	E	F
Topic	CR	Remark	Tweet and Users		
racism	52.4590164	controversial	{'list': [{'text': 'At least the Millwall coac		
hatred	40	controversial	{'list': [{'text': 'if you had allegations in tl		
war	52.7272727	controversial	{'list': [{'text': 'RT @PattyArquette: Noth		
Trump	43.5897436	controversial	{'list': [{'text': 'RT @RandyRainbow: Onc		
soccer	45.4545455	controversial	{'list': [{'text': 'RT @donovanxramsey: Pr		

CSV Result 1

Above are the topics shortlisted from a list of topics mentioned already which is plotted on scatter plot & noted in CSV file. All the topics have CR >= 40 percent. The tweets related to every topic in a list get fetched by API call and the analysis on them is performed using Bag of Word models for positive and negative

sentiment analysis after which they are put into two separate lists i.e., Positive and Negative.

Secondly, an analysis is performed on the tweets present in the Negative List.

For every topic, I am creating a separate Positive and Negative list of tweets to maintain unique values.

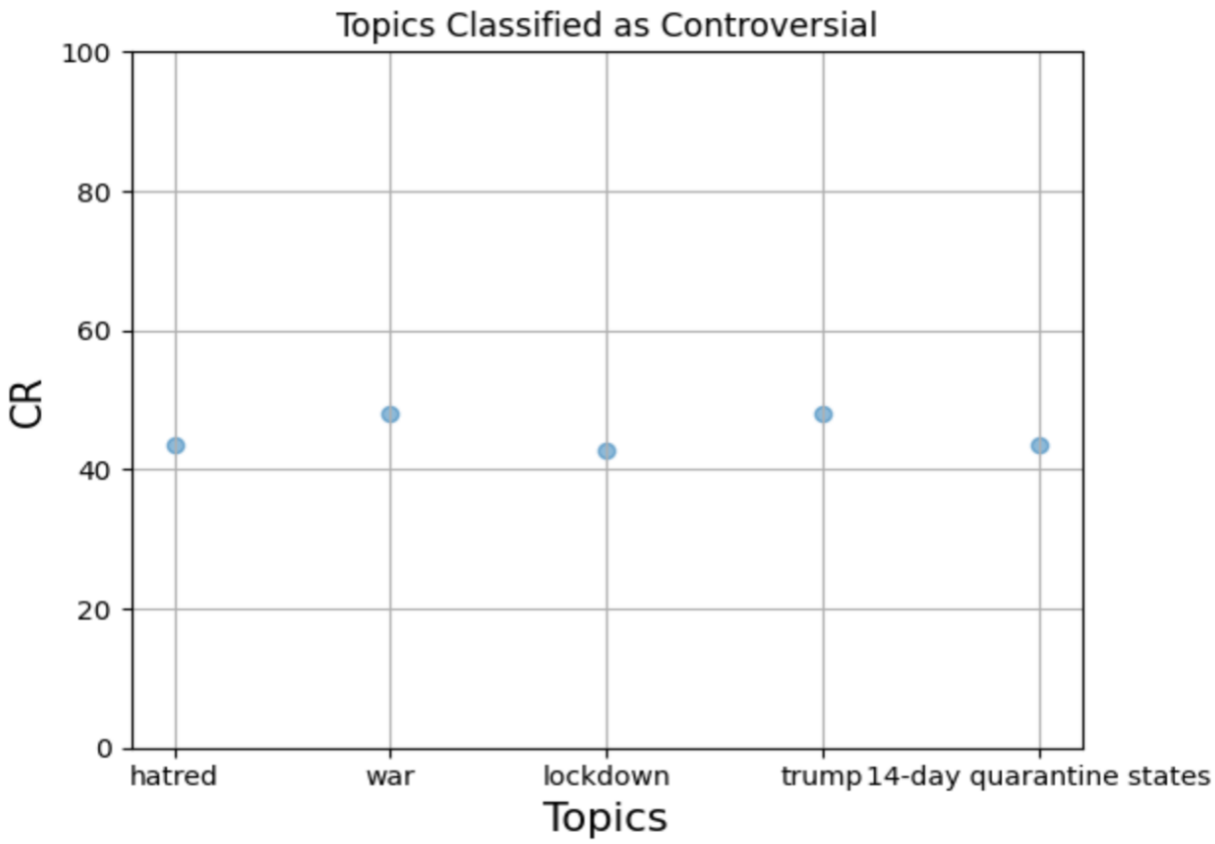
Firstly, I am calculating Total Attention Count (TAC) and Secondly, I am counting Controversial Rate (CR).

Details of how I am calculating them is discussed in detail above at Step 3.

Please note each time the model runs the topics classified as controversial will change based on the current number of tweets fetched and classified as positive and negative.

It fulfills one of the important requirements which this model needs to deal with and i.e., real-time data to monitor real-time trending topics and flagged them as controversial or non-controversial based on the output given by the model.

The accuracy of the model is greatly affected by the limited number of topics my model fetch per API request which is not more than 200. For a real model to work I need tweets but in a large number based on every trending topic.



Plot 1

Topic	CR	Remark	Tweet and Users
hatred	43.54838709677419	controversial	{'list': [{'text': "RT @KyleKulinski: It's funny bc I get shit from people to my left because I firmly believe reform > revolution. But I've always been clear...", 'uid': 2555433834, 'date'
war	47.91666666666667	controversial	{'list': [{'text': '#NowPlaying on RADIO COCCINELLE la radio OVNI Mirage - No more no war -1985 No more no war -1985 Mirage Mirage - N... https://t.co/XJBhbYYUlw', 'uid': 737227
lockdown	42.857142857142854	controversial	{'list': [{'text': 'RT @MarinaDiamandis: Lockdown has basically consisted of me putting on A LOT of amateur cabaret performances for my cat. \n\nHe hates them al...', 'uid': 31485542
trump	48.148148148148145	controversial	{'list': [{'text': "@StephenKing Trump is an evil genius. \nHe's tapped into the known hate that has been around since the civil war bu... https://t.co/jRiEpvn0dC", 'uid': 104009769302
14-day quarantine states	43.47826086956522	controversial	{'list': [{'text': '@CTVNews Don't give a crap about anyone else but themselves. And considering Florida is one of the bigger Covid hot... https://t.co/7yY3ok3FoT', 'uid': 40817605, 'dat

CSV Result 1

When I ran the model second time, I got the above results based on latest tweets at that time on those topics which I already listed as current trending topics

Note-

I tried to add images as clear as possible by zooming in they can be seen clearer.

Bag of Word Model:

The bag-of-words model is a simplifying representation used in natural language processing and information retrieval (IR).

In this model, a text (such as a sentence or a document) is represented as the bag (multiset) of its words, disregarding grammar and even word order but keeping multiplicity. *(Source: Wikipedia)*

Steps to build Bag of Word model.

I created dictionaries for negative contractions and sad smiles. So that I can sanitize each and every tweet and remove links and any other unnecessary stuff.

```
def load_dict_contractions():
    return {
        "ain't": "is not",
        "amn't": "am not",
        "aren't": "are not",
        "can't": "cannot",
        "'cause": "because",
        "couldn't": "could not",
        "couldn't've": "could not have",
        "could've": "could have",
        "daren't": "dare not",
        "daresn't": "dare not",
        "dasn't": "dare not",
        "didn't": "did not",
        "doesn't": "does not",
        "don't": "do not",
        "e'er": "ever",
        "em": "them",
        "everyone's": "everyone is",
        "finna": "fixing to",
        "gimme": "give me",
        "gonna": "going to",
        "gon't": "go not",
        "gotta": "got to",
        "hadn't": "had not",
        "hasn't": "has not",
        "haven't": "have not",
        "he'd": "he would",
        "he'll": "he will",
        "he's": "he is",
        "he've": "he have",
        "how'd": "how would",
        "how'll": "how will",
        "how're": "how are",
        "how's": "how is",
        "I'd": "I would",
        "I'll": "I will",
        "I'm": "I am",
        "I'm'a": "I am about to",
        "I'm'o": "I am going to",
        "isn't": "is not",
```

***Dictionary for contraction of words (total up to 124 commonly used) (Note**: - I only showed a glimpse in above image the actual list is much longer.)*

```
def load_dict_sadsmileys():
    return {
        ":-(": "sad",
        ":-c": "sad",
        ":-<": "sad",
        ":-[": "sad",
        ":(": "sad",
        ":c": "sad",
        ":<": "sad",
        ":[": "sad",
        ":-||": "sad",
        ">:[": "sad",
        ":{": "sad",
        ":@": "sad",
        ">:(": "sad",
        ":'-(": "sad",
        ":'(": "sad"
    }
```

Dictionary for contraction of sad emojis

I used these dictionaries to sanitize each and every tweet before classifying them as Positive and Negative tweet using text blob library.

```
    # create TextBlob object of passed tweet text
    analysis = TextBlob(self.clean_tweet(tweet))
    # set sentiment
    if analysis.sentiment.polarity > 0:
        return 'positive'
    elif analysis.sentiment.polarity == 0:
        return 'neutral'
    else:
        return 'negative'
```

Classification using textblob

Empirical results, including details on the experimental setup and my What were your hypotheses / what do you expect to see from your experiments?

I used Python version 3.7 to build a model for phase 1. I also used libraries such as tweepy, matplotlib, pandas, textblob, CSV, bs4, itertools, and re.

In my algorithm section, I have mentioned all the steps in detail along with diagrams and a flowchart.

The *CR* rate I used is 40 % because of the limitation of Twitter API which gave me only 200 tweets per API request.

Initially, I planned to keep my *CR* rate more than 75 % but my model still works for demo purposes on *CR* = 40 % and gives me a good set of expected results.

After fetching tweets based on the list of topics, I sanitize them removing links, special symbols, etc.

After which I am using *the bag of words* technique to perform *sentiment analysis* of positive and negative tweets per topic by using *text blob* library.

Once they are classified then my model is only concern with negative tweets for the rest of the analysis.

Before moving to the next step model is making sure to calculate *the Total Attention Count (TAC)* for every topic.

TAC is nothing but all the likes, dislikes, positive tweets, and negative comments count for every topic. ***(Inspiration and Source: Controversy trend detection in social media)***

Topics whose *TAC* is higher are getting passed to the next step of analysis first.

I am then calculating *CR* for every topic. CR is equal to the total number of negative tweets, likes, and emojis divided by a total number of positive and negative tweets, likes, and emojis.

Topics whose CR is higher than 40 % are remarked as controversial and saved in the CSV file along with tweets and user lists, remarks, and names of topics.

This CSV file is further used to plot a scatter plot using the matplotlib library.

The one important limitation for this model is the absence of the geolocation of every user who tweeted for a controversial topic.

Phase 2 requires geolocation of users whose name is present inside the CSV file because based on that location the system will alert authorities to take the right action.

That is why I had to keep Phase 2 as a theoretical model and based on assumptions and previous studies I had to formulate reasoning about forecasting the location where unrest may happen based on the topics classified as controversial by my model. I expected to see the topics plotted on scatter plot which are maliciously controversial as well as in CSV file

I expected to see the detailed Tweet and User list, CR value, Topic and its remark which will always be controversial because I am only saving those for simplicity in CSV file.

What worked, what did not, and why?

The classification of topics from maliciously controversial by building a small model in Python worked well in my project. Although there is still a lot of possibility of improvement. For demonstration purposes, I managed to get good relevant results.

Getting the location of users who created such malicious controversies, tweets or retweets is something I am unable to detect because of Twitter API privacy policies and protection rules.

This is why I have to explain Phase 2 theoretically because Phase 2 requires Geolocation of users to perform analysis for forecasting the possibility of whether unrest will occur at a particular location based on user locations, Controversial topics, and Previous crime records on the same locations.

If I manage to get the Geolocation of users, then it's possible to implement Phase 2 of this project and build a sample demo model for demonstration purposes.

Future directions: -

I would like to design and implement a model for Phase 2 also which this time I didn't implement, and I will also do more study on *the COMPAS* algorithm, various forecasting algorithms, as well as try to find a way to get the geolocation of users from Twitter API.

I will also improve my phase 1 implementation and use a more robust approach or Machine learning technique to perform sentiment analysis.

Change in plans due to the limitation of Twitter API: -

At first, I planned to use *the geolocation* of the user which I can get from Twitter API but when I performed my analysis on the data, they were not present, so I had to cancel my Phase 2 model implementation.

The number of tweets per request is not more than 200 because of which I had to lower my CR rate to 40 %.

If you had more time to spend on the project, what would you have liked to do next?

If I had more time, I would study more on topics related to Time Series Forecasting.

Compass algorithm and its implementation in detail

I will improve my Phase 1 implementation and implement techniques such as Latent Semantic Indexing or Term Frequency – Inverse Document Frequency (TF – IDF) which are more suitable in my project instead of the Bag of Words technique to classify Negative and Positive tweets.

I will implement and design a model for Phase 2 which I could not this time due to lack of required features in data.

What advice about the project would you give to future CS 4100 students?

This entire project is exciting as well as there is still a lot of improvement needed to make it more accurate. Phase 2 is still a theoretical model so I will advise them to implement it. Phase 1 still needs improvement for sentiment analysis of tweets as well as more robust techniques can be used to classify topics into controversial or non – controversial such as TF-IDF, Latent Semantic Indexing, HMM, etc.

This project will help them to learn a lot of exciting topics related to Classification techniques, how to use and build models in Python or any programming language of their choice, forecasting algorithms, Event Prediction Methods in Time Series, even probability, and much more.

I will suggest them to read all the topics I mentioned above as well as try to find the bugs and possible improvements in the present model which in return will help this model to become more efficient and robust in giving outputs.

References: -

Jang, Myungha, Shiri Dori-Hacohen, and James Allan. "Modeling controversy within populations." *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*. 2017.

Berk, Richard A., and Susan B. Sorenson. "Algorithmic approach to forecasting rare violent events: An illustration based in intimate partner violence perpetration." *Criminology & Public Policy* 19.1 (2020): 213-233.

Bahrami, Mohsen, et al. "Twitter Reveals: Using Twitter Analytics to Predict Public Protests." *arXiv preprint arXiv:1805.00358* (2018).

Perry, Chris. "Machine learning and conflict prediction: a use case." *Stability: International Journal of Security and Development* 2.3 (2013): 56.

Blair, Robert A., Christopher Blattman, and Alexandra Hartman. "Predicting local violence: Evidence from a panel survey in Liberia." *Journal of Peace Research* 54.2 (2017): 298-312.

Chaiken, Jan, Marcia Chaiken, and William Rhodes. "Predicting violent behavior and classifying violent offenders." *Understanding and preventing violence* 4 (1994): 217-295.

Metz, Cade, and Adam Satariano. "An algorithm that grants freedom, or takes it away." *The New York Times* (2020).

Casselryd, Oskar, and Filip Jansson. "Troll detection with sentiment analysis and nearest neighbour search." (2017).

De La Vega, L. G. M., and V. Ng. "Determining trolling in textual comments." *11th International Conference on Language Resources and Evaluation. Phoenix Seagaia Conference Center Miyazaki, LREC*. 2018.

Sawyer, Rebecca, and Guo-Ming Chen. "The impact of social media on intercultural adaptation." (2012).

Chimmalgi, Rajshekhar Vishwanath. "Controversy trend detection in social media." (2013).

Articles: -

How Social Networking Works
Dave Roos

How Violent Protests Change Politics
Isaac Chotiner

Why Violence Works
Benjamin Ginsberg

Peaceful vs. Non-peaceful protest
Georgia Dover

64% of Americans say social media have a mostly negative effect on the way things are going in the U.S. today
Brooke Auxier

The Dark Psychology of Social Networks
Jonathan Haidt and Tobias Rose-Stockwell