CSC – 501

REPORT

ASSIGNMENT 4: TEXT DATA

SUBMITTED BY: GROUP C

DIVYANSH BHARDWAJ (V00949736)

ARSHIYA GULATI (V00949938)

SHAIMA PATEL (V00949940)

VENISH PATEL (V00949300)

SUBMITTED TO:

Prof. Sean Chester (schester@uvic.ca)

INTRODUCTION

The data provided in this assignment had 13 different .csv files which were read from " https://github.com/fivethirtyeight/russian-troll-tweets". The files contain list of authors who were involved in biasing the 2016 Trump election by spreading fake rumours regarding politics. The corruption of tweets started by Russians in 2012 gradually increased therefore ruining the elections. Although, this took a while to get notice and so a final proof of this scam was announced just before election in October 2016 by US that Russian Internet Research Agency (IRA) also called "troll factory" was involved in corrupting the November 8th 2016 elections. As a proof of this scam Twitter reported a list of all doubtful twitter-user which were involved with IRA to US government. Hence, we got a chance to go through such twitter handlers and come up with different interesting insights that displays the falsification of the presidential election in our 4th assignment.

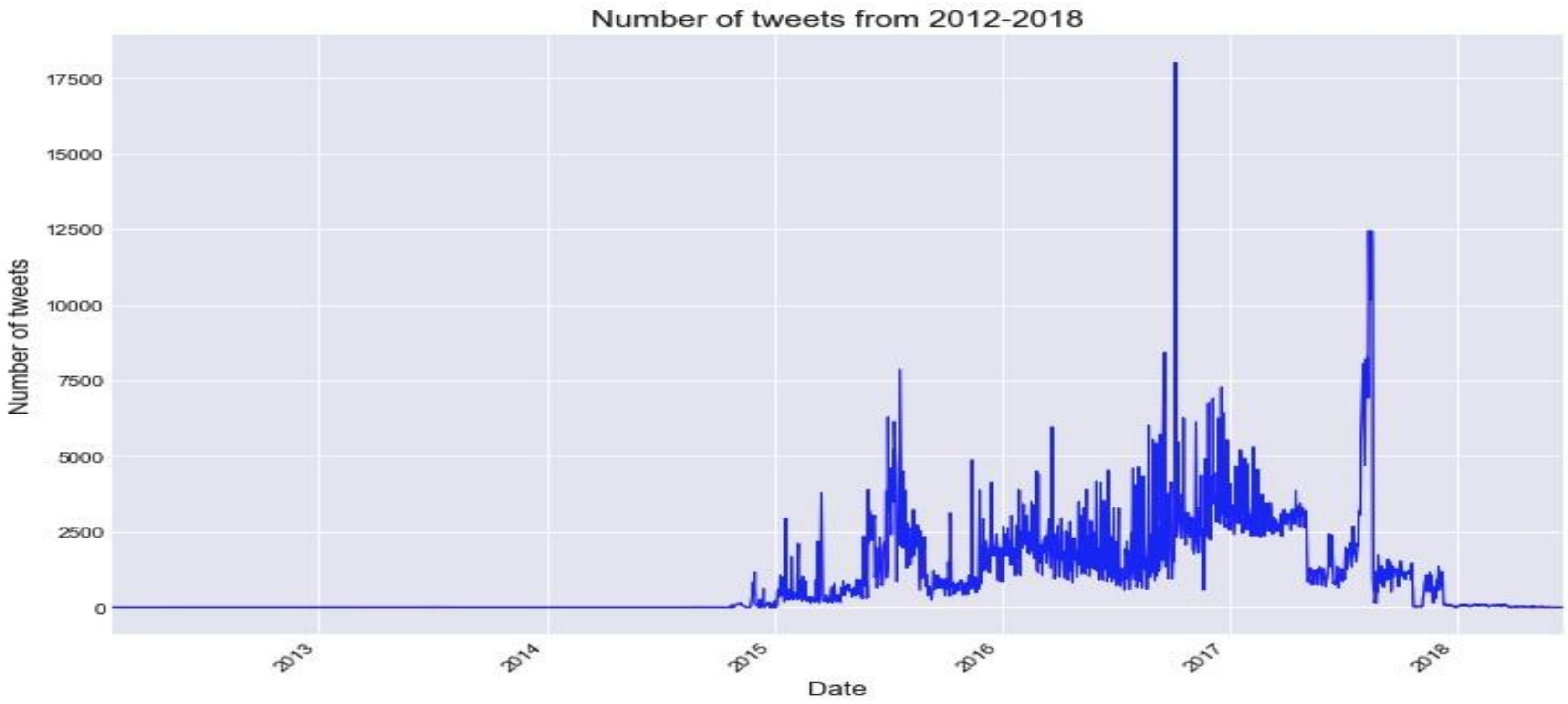
We started this assignment by studying the data provided in .csv files and trying to see the link of what kind of twitter content was involved in discussion related to politics (especially Trump and Hillary). As there were 13 files our initiative was to merge them into one file and see the combined output but as the data processing was a crucial step involving lots of cleaning and appropriate merging, we decided to work with one .csv file and then go for the combined file. We started to model the data using Word2Vec to form vectors of different words. As text modelling was the base of this assignment we decided to give it more importance and use different creative ideas. We used NLTK, Gensim, Word2Vec, SpaCy, Text Blob and KeyedVectors. Further, we proceeded with modelling our data and looking for Biasing. Bias was a very fascinating topic and we explored and learned quite a lot for it. We cleaned the data by removing all Non-English languages, emoji's, non-politic content, punctuations, numbers, typos and combined all the 13 .csv files. After pre-processing the data we modelled the text on our cleaned final .csv   file and formed our final vectors and biases. As this assignment expected an interesting story in form of visualizations, we dedicated most of our times in plotting intriguing insights. We decided to use different columns related to tweets, authors, contents, harvested_date, hashtags etc. It was a very amusing experience as the output of   graphs really got us more and more involved and we plotted more insights to see what actually happened in 2016 election. The story got formed parallel as we proceeded with insights. We also decided to plot sentiment analysis to see the emotions of people related to the election and a very interesting result was obtained. Finally, we read the research papers which comprehended us and linked us to our modelling ideas.

Thus, this assignment gave us a chance to explore different libraries involved with text and also gave an opportunity to have a look at 2016 US elections.

RUBRIC – 1

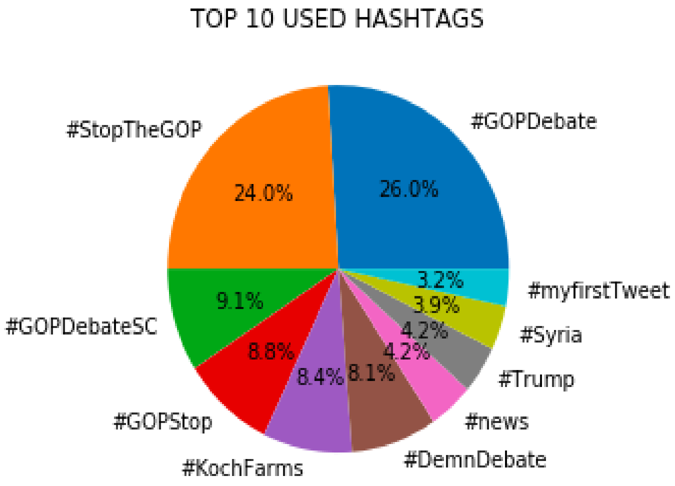
Story Telling

Plot 1: Tweet Count from 2012-2018



The plot above displays number of tweets over 6 years (i.e. 2012-2018). The Russian Troll Factory started corrupting the 2016 election from 2012 but as the graph shows the tweets over 2012-2015 were quite minimal. With end of 2015 the tweets increased and it went through some fluctuations but as the day of election approached tweets went on increasing. 8th November, 2016 which was the day of election, number of tweets reached its peak with almost 1800 tweet count. This shows how major was the involvement of twitter_handlers in scamming the elections. With the end of election, in the year 2017-2018 tweets regarding political content gradually decreased.

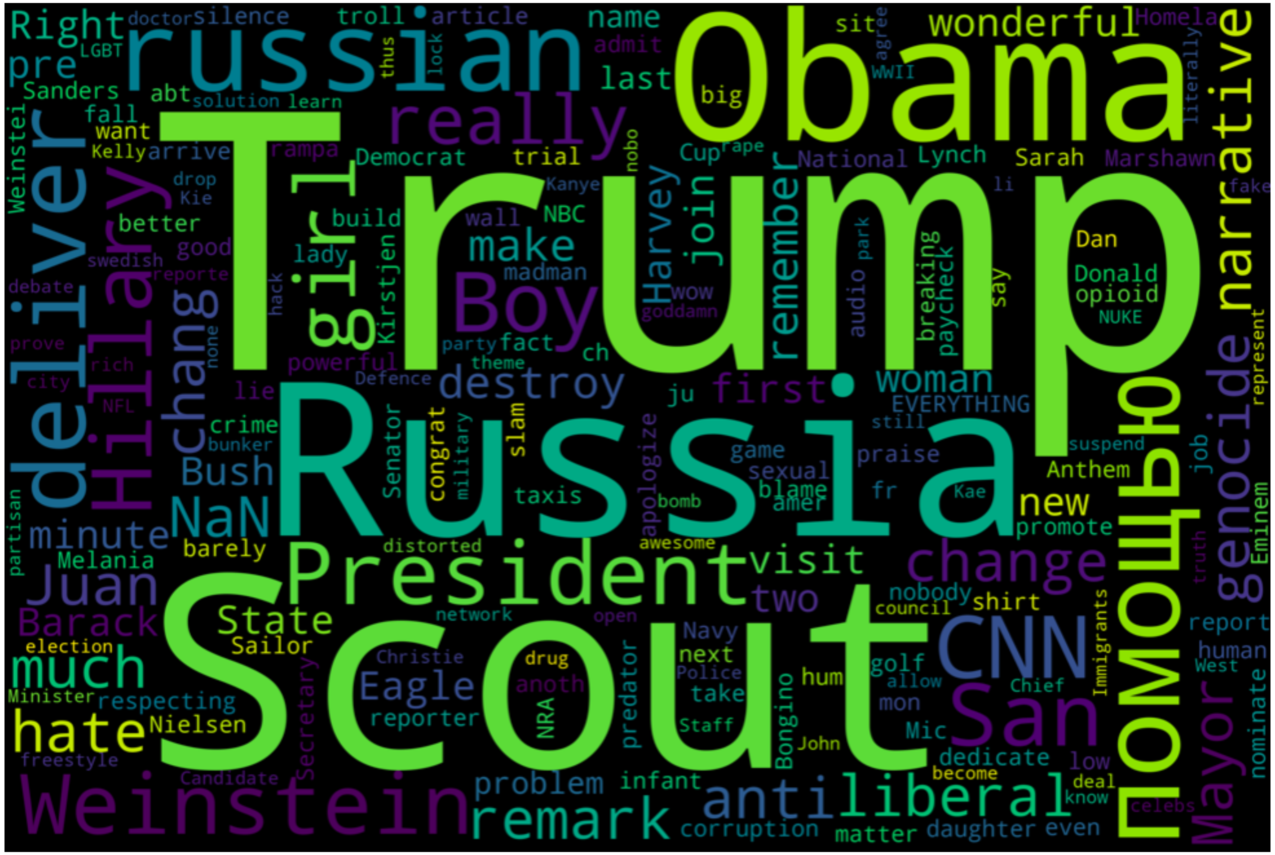
Plot 2: Top 10 Hashtags used in Tweets during Elections



As the tweet_count showed some interesting results, we decided to see which hashtags were most popular over election days. As assumed the hashtags related to elections and political party were seen quite often in the tweets. The hashtag related to Grand Old Party (GOP) i.e. Republican National Committee of US were quite a lot in discussion. These hashtags which were used in tweets almost talked about GOP as the election approached and hence we could prove trolls talked quite a lot about politics in their tweets and used many politics related hashtags.

Plot 3: Words with maximum occurrence in tweets

The below depicted word cloud shows which words were maximum used in tweet content by the trolls. We decided to form the word cloud as we were interested in seeing what specific type of discussion related to elections were happening on tweets and hashtags. The name of political leaders, political parties, Russia, Scouts (to whom Trump addressed a speech after election) remained to be the most used words over election days. This shows that the tweets and tags contained words related to elections and especially Trump.


Trump
Scout
Russia
Obama
russian
President
Weinstein
deliver
Hillary
помощью
narrative
genocide
Boy
really
girl
CNN
San
antiliberal
remark
change
Mayor
Juan
NaN
Bush
chang
hate
much
Barack
minute
destroy
first
woman
remember
join
Harvey
make
visit
two
State
Eagle
new
wonderful
Right
pre
name
last
troll
article
silence
doctor
LGBT
Sanders
fall
want
abt
solution
learn
Kelly
arrive
rampa
Democrat
trial
Cup
rape
National
Lynch
Sarah
Marshawn
Homela
sit
agree
WWII
big
literally
fake
li
Kanye
nobo
wall
NBC
build
better
drop
Kie
good
lady
madman
Kirstjen
swedish
reporte
debate
hack
none
prove
rich
city
NFL
powerful
Defence
party
fact
theme
goddamn
wow
ch
audio
park
breaking
paycheck
say
Dan
Donald
opioid
NUKE
represent
lie
ju
EVERYTHING
still
suspend
job
crime
bunker
Senator
congrat
military
slam
apologize
game
sexual
blame
amer
bomb
fr
praise
Kae
Anthem
Eminem
taxis
promote
partisan
Melania
barely
distorted
awesome
truth
nobody
council
shirt
Immigrants
network
open
report
human
West
nominate
election
Sailor
Christie
drug
next
Police
golf
allow
Chief
Minister
respecting
anoth
predator
hum
take
mon
Nielsen
reporter
NRA
Staff
Bongino
Mic
freestyle
Secretary
John
dedicate
low
Candidate
problem
infant
become
deal
know
celebs
corruption
matter
daughter
even
Weinstei
thus
lock

Plot 4: Sentiment Analysis Before, On and After Election

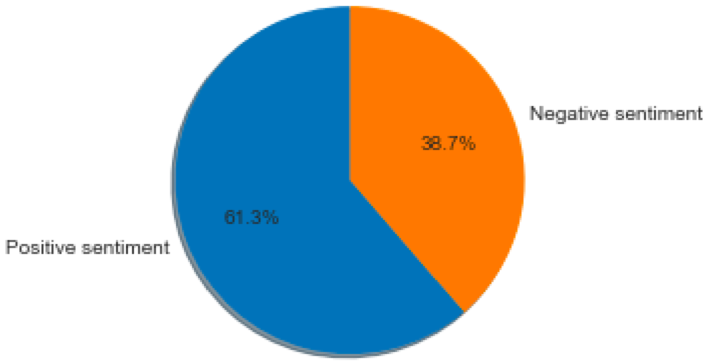


Figure 1 Sentiments Before Election

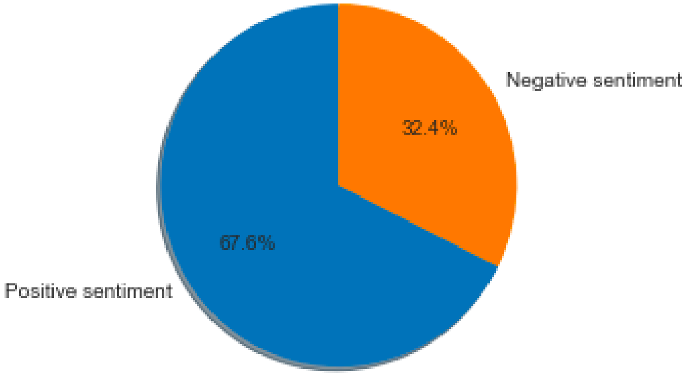


Figure 2 Sentiments On Election

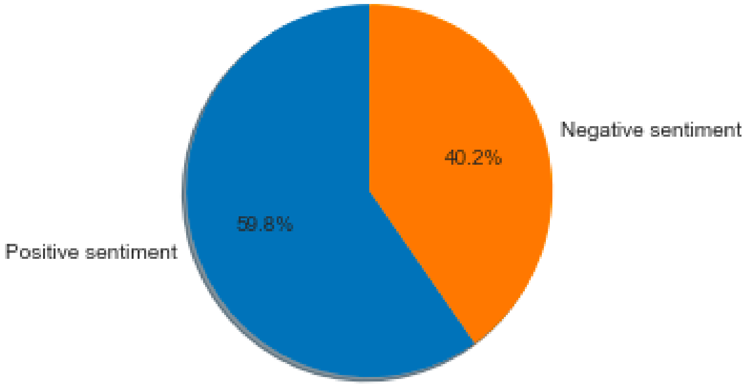
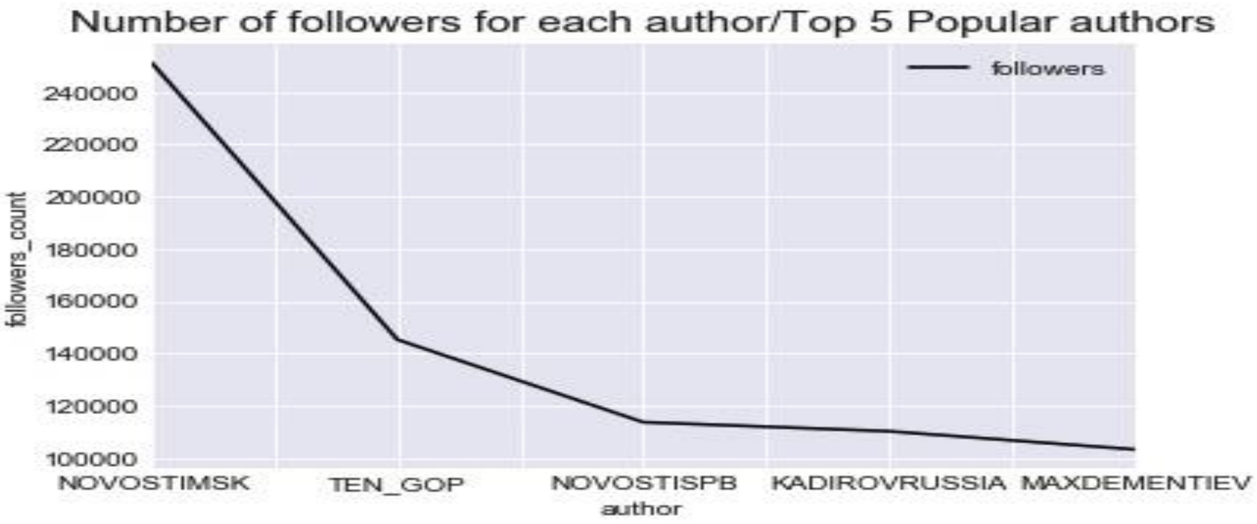


Figure 3 Sentiments After Election

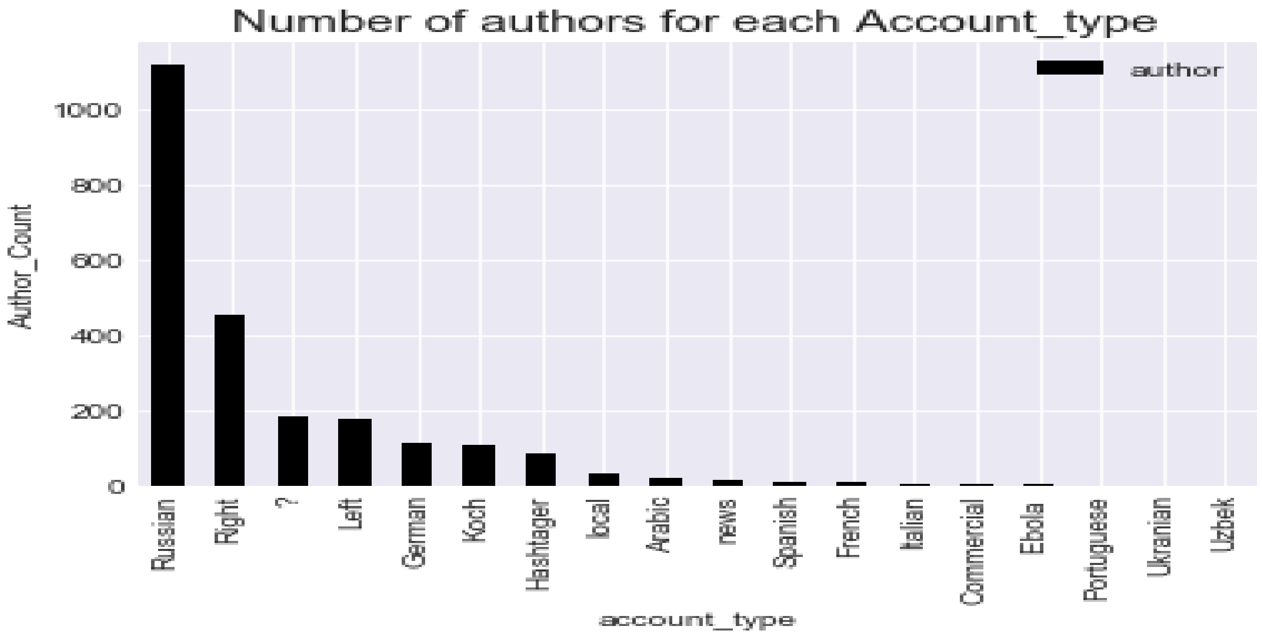
As the plot-3 showed that election was the talk around town so we decided to plot sentiment analysis to see what were people's emotion towards it. We added a column in our .csv file calculating the polarity of words used in a given sentence. +1 = Positive Sentiment and -1 = Negative Sentiment. We found that 38.7% people (fig-1) had negative views towards elections which decreased to around 32 % on election day (fig-2). But after election (fig-3) the negative sentiment increased and positive sentiment decreased. This shows that the Russians way of showing negativity towards election succeeded and so views of people towards 2016 election became negative.

Plot 5: Authors with Maximum Followers



From this visualisation we get to know the most popular authors i.e. the authors that have maximum number of followers and they are NOVOSTIMSK and TEN _GOP. These were actually the accounts that were being handled by the Russian government. It was done to control the elections in the west. Many of these accounts have been suspended from the twitter such as TEN_GOP. This shows that negativity increased after election was due to maximum followers being of Russian Authors who intended to spread negativity.

Plot 6: Account_type with maximum authors



This visualisation plots the account type which has maximum authors. As we saw from plot – 5, Russian Authors had most number of followers and plot-6 shows that there were maximum Russian Accounts which confirms that the election was corrupted by Russian IRA. Hence, this proves that Russian trolls were involved in biasing the 2016 Presidential Election.

RUBRIC – 2

Data Modelling

We trained the given data into model using the concept of word embeddings such as word2vec and bag of words. We designed a bigram as well. And also trained the model after applying tokenization. Where tokenization is converting a sentence into words which are further modelled into a vector by word2vec.

We preferred word2vec over bag of words as in this context information is not lost and that the low dimension vectors are formed.

The modelled data is seen to be biased and this trained model is said to be a corpus.

The main benefit of our corpus-specific embeddings is the presented word cloud given by us is only specifically to the results of the election while the cloud formed by the google pre trained data is giving different visualization which is not related to the elections.

We also took the pre trained model into consideration.

Advantage

The major benefit of using this pre-trained data is that it has more vocabulary than the other one so it is easy to perform lemmatization and other such data clearing aspects.

Whereas advantage of using corpus specific model which we trained is that it completely relates to our dataset.

Disadvantage

When we applied the pre trained data it was not giving the solutions according to the expected results as it is having lots of data not connected and modelled according to our data of elections.

Whereas corpus specific model gives more efficient results, relating to our dataset.

We have applied and want to give example of biasing to differentiate the results of pre-trained and corpus-specific embeddings.

But before modelling the data we cleaned the data removing the urls, hashtags, handles and took only tweets relating to the elections and its political parties into considerations.

We have also done the process of lemmatization which means removing inflectional endings from the word and giving the most basic form of word.

The cleaning is done to make our data more efficient and make our visualisations more clear.

Pre-trained	girl	President	'boy', 0.5697915554046631
	boy	leader	'girl', 0.6256637573242188
	election	Russia	'elections', 0.6523491144180298

Corpus-specific	girl	President	'Sad', 0.515403151512146
	boy	leader	'girl', 0.6052294969558716
	election	Russia	'collusion', 0.5787917971611023

RUBRIC – 3

Reproducibility

The pre-processing of data for this assignment was quite crucial to form a meaningful and insightful story. We merged the 13 .csv files and cleaned them to form a proper representation.

COMBINING .CSV FILES:

- We used OS and Glob libraries to combine 13 .csv’s.
- Firstly, we provided path to 13 .csv files and then we combined them into 1 single file.
- We read the files and concatenated it followed by storing it in a dataframe.
- Finally, we converted the output into a single .csv file.

CLEANING THE FINAL .CSV FILE:

- We used gensim, nltk and spaCy library over our merged .csv file.
- After merging, we separated English language content and removed all Non-English contents.
- We then also removed urls, hashtags and handles.
- After that we extracted each token by lemmatizing it.
- We also removed characters with less than 3 alphabets and omitted stopwords.
- We removed all useless stuff like space, numbers, emojis etc.
- Finally, we removed all tweets with non-political content and kept only politics related tweets.
- The file after cleaning had 2116866 rows which were originally 2946207. This shows that a significant amount of cleaning was done.

After merging and cleaning the csv file we implemented visualizations. Talking about reproducibility we can clearly see that the visualization properly demonstrates the dataset.

RUBRIC – 4

Connection To Research

We faced many challenges for removing stop words, hashes and also for pre-processing data. We also needed to know exactly how to pre train data which was then done by going through research paper. It helped us cleaning data and get the output we actually wanted to work on.

After that we started reading the research paper named "Mitigating Gender Bias in Natural Language Processing". It described how there are different forms of bias and how can we use biasing to see and test our data . Hence we used the research paper as our base and started forming a bias on our pre-trained corpus. To see if the our pre-trained data is bias or not we had to compare it with another pre-trained data which we took from google. Comparing these made us understand what bias is and how does it really look like. The research paper was quite helpful to understand the theoretical concept and using it when we plotted graph we could easily prove that corpus which we trained was bias. We tested some of the gender related words into positive and negative values.

Eg. Positive = ['boy','leader'] gave result = [('girl', 0.6256637573242188)]

This proves that our data is bias and so we tried doing this using many different words to confirm it.

"GoogleNews-vectors-negative300.bin.gz"- Link we used for google imported pre-trained data

The other research paper also helped us to know what different kind of visualizations we have in text data. We got to know that text does not have a normal visualization rather it uses enhance version known as word cloud. We tried to implement words in our tweets to form word cloud by using research paper and got a satisfactory output.