INSY 5378 DATA SCIENCE PROGRAMMING APPROACH

GROUP 7

Social Media Analytics

**Team Members:**

Salman Mohammed-1001398053 [salman.mohammed@mavs.uta.edu](mailto:salman.mohammed@mavs.uta.edu) Information Systems

Sujatha Sivakumar-1001359237 [sujatha.sivakumar@mavs.uta.edu](mailto:sujatha.sivakumar@mavs.uta.edu) Information Systems

Tapan Patel-1001450804 [tapansaileshbha.patel@mavs.uta.edu](mailto:tapansaileshbha.patel@mavs.uta.edu) Information Systems

Table of Contents .

**Introduction …………………………………………………………………………………….. 3**

**Twitter Streaming ……………………………………………………………………………. 3**

**Pre-Processing ………………………………………………………………………………… 3**

**Sentiment Analysis …………………………………………………………………………. 3**

**Word Cloud ……………………………………………………………………………………. 6**

**Topic Modeling ………………………………………………………………………………. 9**

**Insights …………………………………………………………………………………………… 15**

# Introduction

Donald J. Trump became the 45th President of the United States on January 20, 2017. Since the inauguration, President Trump is actively initiating new policies and conversations, which generate active conversations in the social media.

In this project we collected all such generated traffic related to President Trump from twitter and performed social media analytics to understand the opinion among the people regarding the president.

# Twitter Streaming

We have used twitter streaming API to collect tweets from twitter. Using the credentials obtained from the API we have used Tweepy module to collect real time tweets from all over the USA using hash tags like trump ,Donald trump, POTUS. We collected 10,000 such tweets.

We have also collected demographic specific tweets using geo location co-ordinates for five different states- California, Michigan, Kansas, Texas, Pennsylvania. We collected 1000 tweets from each location.

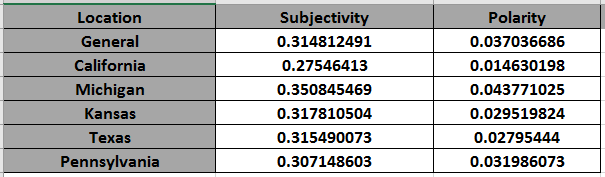
# Pre-Processing

We extracted the tweets text from the raw JSON format obtained from the streaming API and cleansed the text to separate URLs, punctuations, digits, emoticons and Unicode using HTMLParser, Preprocessor, RegEx modules. The preprocessed text was then stemmed and lemmatized using Lancaster stemming. The Stemmed text was then cleansed of stop words using the NLTK module. Also, special words such as “Trump”, “Donald” etc. were removed from the text.

# Sentiment Analysis

We analyzed the sentiments of each tweet to calculate the subjectivity and polarity scores of the tweets. Here subjectivity means that there is a context in the text. High subjectivity scores mean that the tweets have proper context. Polarity on the other hand portrays the nature of the content. The value varies from -1 to +1 where minus values indicate that the tweets are negative and positive values indicate that the tweets are positive.

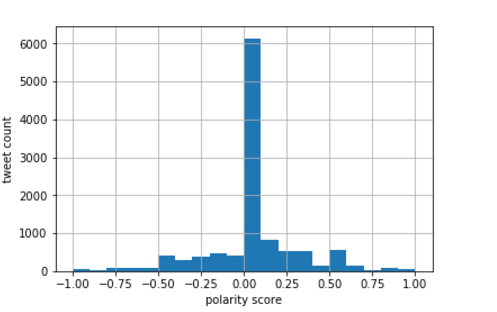
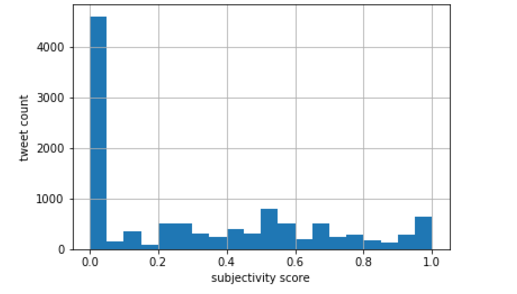
We used text blob to calculate the average subjectivity and polarity scores. For the tweet corpus. The same was repeated for the demographics specific tweets.



**The Graphs are as follows:**

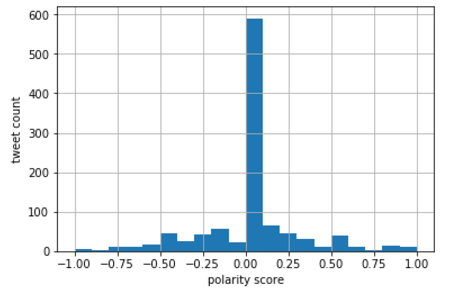
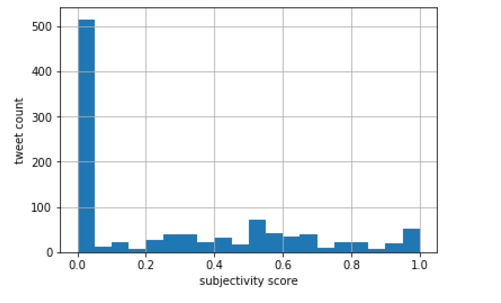
**Trump General:**

**Subjectivity: Polarity:**



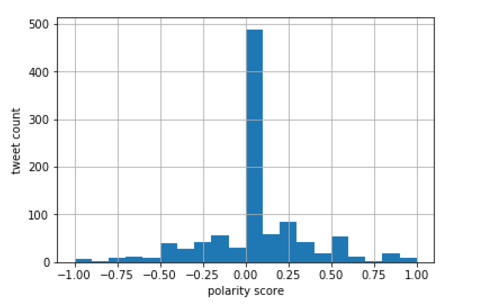
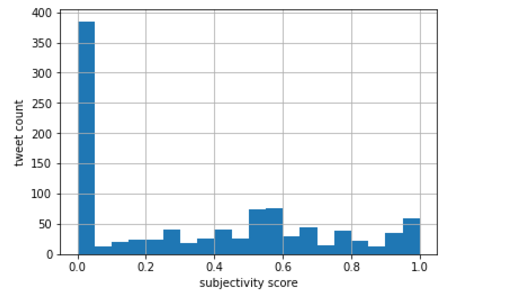
**California:**

**Subjectivity: Polarity:**



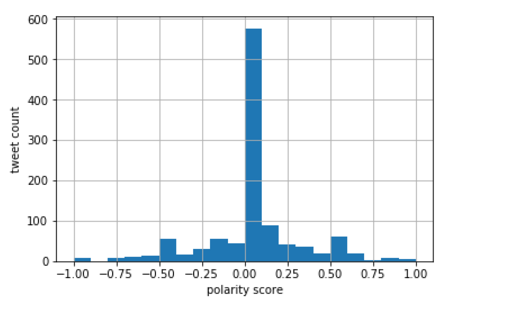
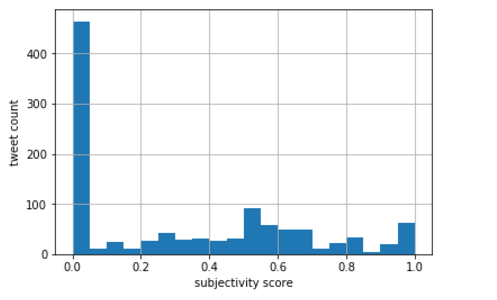
**Michigan:**

**Subjectivity: Polarity:**



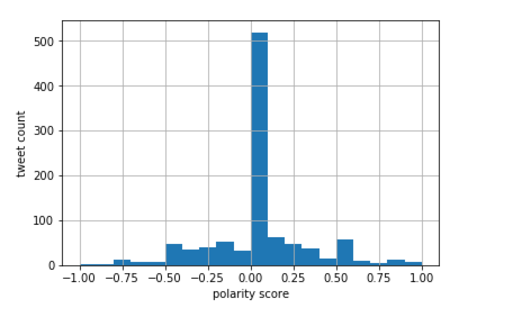
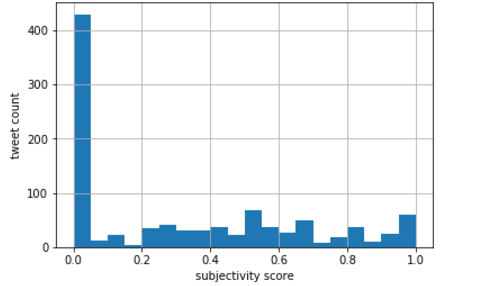
**Kansas:**

**Subjectivity: Polarity:**



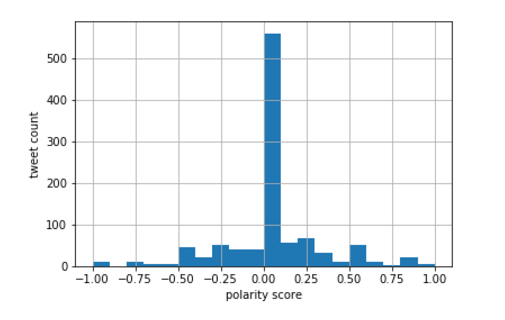
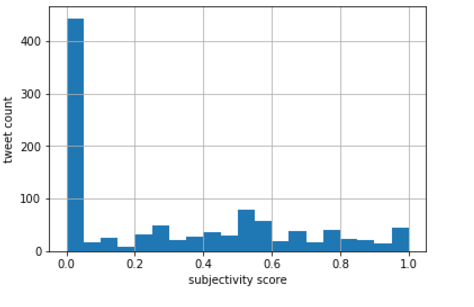
**Texas:**

**Subjectivity: Polarity:**



**Pennsylvania:**

**Subjectivity: Polarity:**

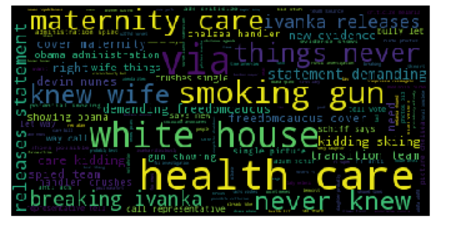


From the above results, we find that there is maximum subjectivity and polarity score in Michigan. This means that people are really being positive about the new president in Michigan. On a general note, we find that there is nearly positive comments about the president.

# Word Cloud Generation

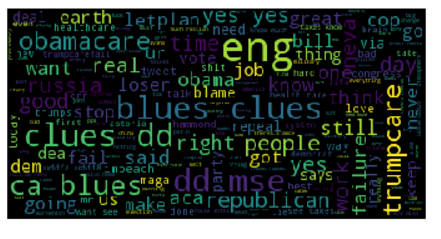
We generated a word cloud from the cleansed text using word cloud module to find out the most common words used in the tweets about the president with the data from all over the USA as well as from different states.

General



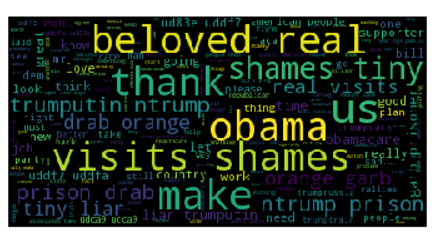
From the figure we can see that the most used words were Maternity care, Smoking Gun, White House, Health Care and about the president’s family.

**California**



From the figure we can see that the most used words were trump care, Obama care.

**Michigan**



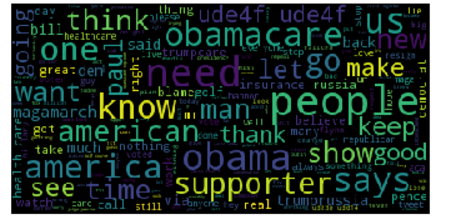
From the figure we can see that the most used words were Obama, visits, shames, beloved.

**Kansas**



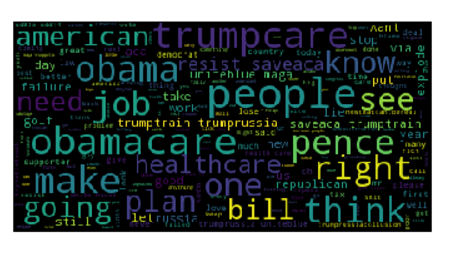
From the figure we can see that the most used words were Russia, Obama care, paul.

**Texas**



From the figure we can see that the most used words were Obama care, supporter, America, Obama.

**Pennsylvania**



From the figure we can see that the most used words were job, bill , trump care, Obama care, republican.

# Topic Modelling

From the cleansed text and word cloud we created two models for analyzing the trending topics in the tweets. We used two machine learning algorithms to do this - the Non- Negative Matrix Factorization & Latent dirilect Allocation . The results are as follows.

**General:**

**NMF:**

[4432 4504 5079 2923 3811 2293 1762]

russia

says

team

like

people

health

evidence

[3603 4504 5079 2923 2293 3811 1762]

obama

says

team

like

health

people

evidence

[ 774 3811 1762 756 5175 6084 656]

care

people

evidence

campaign

time

white

breaking

[2397 5079 1762 756 5175 656 6084]

house

team

evidence

campaign

time

breaking

white

[3409 1762 756 5175 6084 656 1878]

new

evidence

campaign

time

white

breaking

fbi

[5974 4504 2923 5079 2293 3811 2616]

vote

says

like

team

health

people

investigation

[3573 4504 5079 2293 3811 1762 2616]

nunes

says

team

health

people

evidence

investigation

/\*\*//\*\*//\*\*//\*\*//\*\*//\*\*//\*\*//\*\*//\*\*//\*\*/

(94911, 6228)

5 281.1761014031083

10 277.63825384985944

15 274.8344958639479

20 272.45733636863696

25 270.255435173741

**LDA:**

Topic #1 (0, '0.021\*"nunes" + 0.017\*"voters" + 0.017\*"utah" + 0.016\*"health" + 0.013\*"trumprussia" + 0.011\*"us" + 0.010\*"things" + 0.009\*"breaking" + 0.009\*"cnn" + 0.009\*"bid"')

Topic #2 (1, '0.032\*"russia" + 0.020\*"says" + 0.018\*"new" + 0.017\*"via" + 0.016\*"evidence" + 0.013\*"would" + 0.013\*"bill" + 0.012\*"putin" + 0.010\*"russian" + 0.009\*"know"')

Topic #3 (2, '0.032\*"obama" + 0.025\*"care" + 0.022\*"house" + 0.016\*"investigation" + 0.015\*"photos" + 0.015\*"ads" + 0.012\*"right" + 0.011\*"criticize" + 0.010\*"surveillance" + 0.009\*"knew"')

Topic #4 (3, '0.020\*"anti" + 0.017\*"like" + 0.016\*"people" + 0.016\*"melania" + 0.013\*"never" + 0.012\*"said" + 0.011\*"de" + 0.010\*"best" + 0.010\*"call" + 0.008\*"way"')

Topic #5 (4, '0.018\*"vote" + 0.016\*"team" + 0.016\*"rt" + 0.016\*"news" + 0.014\*"campaign" + 0.014\*"time" + 0.013\*"fbi" + 0.013\*"get" + 0.011\*"white" + 0.011\*"see"')

**California:**

**NMF:**

[244 555 967 317 329 328 327]

eng

neva

yuge

gets

gonzaga

gonna

gone

[964 920 377 127 550 575 608]

yes

want

house

care

need

obamacare

people

[196 163 818 967 328 327 326]

dd

cop

today

yuge

gonna

gone

golfing

[103 358 842 722 807 688 558]

blues

healthcare

trumps

says

thing

republicans

news

[144 555 672 209 597 806 321]

clues

neva

real

dems

party

theresistance

gives

[121 705 127 608 833 864 193]

ca

russia

care

people

trumpcare

ud83e

day

[540 575 440 815 864 713 946]

mse

obamacare

know

time

ud83e

said

work

/\*\*//\*\*//\*\*//\*\*//\*\*//\*\*//\*\*//\*\*//\*\*//\*\*/

(7028, 968)

5 65.6886597201659

10 63.749510022670194

15 62.57795348375914

20 61.58733663996545

25 60.76183289876763

**LDA:**

Topic #1 (0, '0.031\*"yes" + 0.016\*"good" + 0.013\*"want" + 0.011\*"repeal" + 0.011\*"dems" + 0.010\*"healthcare" + 0.009\*"mr" + 0.009\*"right" + 0.008\*"congress" + 0.008\*"money"')

Topic #2 (1, '0.070\*"eng" + 0.031\*"trumpcare" + 0.016\*"dd" + 0.013\*"russia" + 0.013\*"fail" + 0.013\*"obama" + 0.010\*"job" + 0.010\*"plan" + 0.009\*"deal" + 0.009\*"take"')

Topic #3 (2, '0.020\*"obamacare" + 0.019\*"know" + 0.016\*"ca" + 0.014\*"clues" + 0.010\*"got" + 0.009\*"blues" + 0.008\*"failure" + 0.008\*"real" + 0.007\*"think" + 0.007\*"health"')

Topic #4 (3, '0.038\*"get" + 0.029\*"like" + 0.027\*"people" + 0.017\*"bill" + 0.017\*"mse" + 0.016\*"neva" + 0.013\*"care" + 0.011\*"let" + 0.011\*"house" + 0.010\*"world"')

Topic #5 (4, '0.014\*"ude06" + 0.012\*"time" + 0.011\*"would" + 0.010\*"work" + 0.008\*"stop" + 0.007\*"says" + 0.007\*"day" + 0.007\*"make" + 0.007\*"office" + 0.007\*"ur"')

**Michigan:**

**NMF:**

[ 671 631 603 959 830 1014 330]

people

obama

need

trumputin

shames

visits

garb

[748 631 649 603 929 234 72]

real

obama

orange

need

tiny

drab

beloved

[477 631 625 649 929 72 234]

know

obama

ntrump

orange

tiny

beloved

drab

[ 509 631 929 72 234 1014 718]

like

obama

tiny

beloved

drab

visits

prisoner

[500 603 234 929 959 72 718]

liar

need

drab

tiny

trumputin

beloved

prisoner

[981 631 929 330 718 72 540]

ud83e

obama

tiny

garb

prisoner

beloved

maga

[ 717 625 649 603 1014 959 234]

prison

ntrump

orange

need

visits

trumputin

drab

/\*\*//\*\*//\*\*//\*\*//\*\*//\*\*//\*\*//\*\*//\*\*//\*\*/

(7901, 1086)

5 71.1969100461767

10 69.89277515270108

15 68.68042059592737

20 67.49074063475594

25 66.43041861139926

**LDA:**

Topic #1 (0, '0.022\*"ntrump" + 0.021\*"orange" + 0.021\*"american" + 0.014\*"care" + 0.010\*"even" + 0.010\*"go" + 0.009\*"want" + 0.009\*"repeal" + 0.009\*"pres" + 0.009\*"sad"')

Topic #2 (1, '0.031\*"people" + 0.028\*"know" + 0.021\*"liar" + 0.016\*"like" + 0.015\*"job" + 0.015\*"great" + 0.014\*"ties" + 0.014\*"bill" + 0.013\*"obamacare" + 0.012\*"would"')

Topic #3 (2, '0.023\*"real" + 0.017\*"russia" + 0.016\*"us" + 0.015\*"get" + 0.014\*"drab" + 0.013\*"shames" + 0.013\*"visits" + 0.013\*"tiny" + 0.012\*"really" + 0.012\*"garb"')

Topic #4 (3, '0.022\*"prison" + 0.014\*"one" + 0.013\*"work" + 0.013\*"time" + 0.012\*"trumpcare" + 0.010\*"still" + 0.009\*"golfing" + 0.008\*"today" + 0.008\*"big" + 0.008\*"course"')

Topic #5 (4, '0.018\*"need" + 0.013\*"obama" + 0.011\*"maga" + 0.010\*"let" + 0.009\*"open" + 0.008\*"public" + 0.008\*"associates" + 0.008\*"think" + 0.007\*"hearings" + 0.007\*"man"')

**Kansas:**

**NMF:**

[ 684 713 870 549 836 1219 1430]

like

maga

paul

house

obamacare

time

white

[ 487 367 713 549 717 836 1219]

golf

en

maga

house

make

obamacare

time

[ 835 367 713 870 1219 717 836]

obama

en

maga

paul

time

make

obamacare

[ 490 870 713 1219 359 809 70]

good

paul

maga

time

el

news

america

[1036 713 870 359 519 1430 70]

russia

maga

paul

el

health

white

america

[ 636 367 870 359 809 808 1400]

know

en

paul

el

news

new

voted

[ 880 1205 367 717 549 1219 809]

people

think

en

make

house

time

news

/\*\*//\*\*//\*\*//\*\*//\*\*//\*\*//\*\*//\*\*//\*\*//\*\*/

(9332, 1471)

5 80.21845174450405

10 79.28429881508596

15 78.39643003656813

20 77.60154768408847

25 76.85701134407385

**LDA:**

Topic #1 (0, '0.016\*"tiempos" + 0.015\*"amor" + 0.011\*"obamacare" + 0.010\*"health" + 0.009\*"make" + 0.007\*"people" + 0.007\*"putin" + 0.006\*"show" + 0.005\*"u2705trump" + 0.005\*"man"')

Topic #2 (1, '0.031\*"de" + 0.009\*"news" + 0.009\*"voted" + 0.008\*"la" + 0.007\*"new" + 0.007\*"antifa" + 0.007\*"rallies" + 0.007\*"world" + 0.006\*"let" + 0.006\*"boris"')

Topic #3 (2, '0.017\*"obama" + 0.016\*"golf" + 0.012\*"america" + 0.011\*"watch" + 0.011\*"would" + 0.011\*"deal" + 0.011\*"bill" + 0.010\*"supporters" + 0.010\*"healthcare" + 0.008\*"tv"')

Topic #4 (3, '0.032\*"en" + 0.031\*"el" + 0.016\*"like" + 0.016\*"us" + 0.013\*"get" + 0.013\*"care" + 0.013\*"one" + 0.012\*"u0131" + 0.010\*"tweets" + 0.010\*"white"')

Topic #5 (4, '0.016\*"russia" + 0.015\*"maga" + 0.013\*"house" + 0.010\*"epshteyn" + 0.009\*"said" + 0.009\*"cia" + 0.008\*"think" + 0.008\*"war" + 0.008\*"husband" + 0.007\*"says"')

**Texas:**

**NMF:**

[ 546 648 511 588 380 539 1042]

like

news

know

man

golf

let

ude4f

[712 588 948 380 956 583 379]

people

man

think

golf

time

magamarch

going

[ 677 648 511 708 1042 539 983]

obamacare

news

know

paul

ude4f

let

trumpcare

[676 588 380 948 917 379 583]

obama

man

golf

think

supporters

going

magamarch

[ 153 648 708 1042 379 1084 585]

care

news

paul

ude4f

going

watch

make

[ 54 588 917 583 1084 1023 390]

america

man

supporters

magamarch

watch

ud83e

great

[834 642 588 648 539 708 917]

says

need

man

news

let

paul

supporters

/\*\*//\*\*//\*\*//\*\*//\*\*//\*\*//\*\*//\*\*//\*\*//\*\*/

(7467, 1136)

5 68.3885955409885

10 67.34240867838243

15 66.46803765166341

20 65.6734352051827

25 64.9230317028779

**LDA:**

Topic #1 (0, '0.024\*"us" + 0.016\*"would" + 0.016\*"insurance" + 0.011\*"know" + 0.011\*"ur" + 0.010\*"jeanine" + 0.010\*"paul" + 0.010\*"well" + 0.010\*"stop" + 0.010\*"work"')

Topic #2 (1, '0.033\*"like" + 0.015\*"america" + 0.013\*"going" + 0.011\*"good" + 0.011\*"want" + 0.010\*"back" + 0.010\*"nothing" + 0.010\*"see" + 0.009\*"let" + 0.008\*"show"')

Topic #3 (2, '0.018\*"care" + 0.017\*"says" + 0.015\*"obamacare" + 0.012\*"keep" + 0.010\*"bannon" + 0.009\*"could" + 0.009\*"health" + 0.008\*"read" + 0.008\*"much" + 0.008\*"stand"')

Topic #4 (3, '0.032\*"people" + 0.031\*"get" + 0.018\*"need" + 0.013\*"trumpcare" + 0.011\*"thanks" + 0.011\*"believe" + 0.011\*"via" + 0.010\*"still" + 0.010\*"anti" + 0.009\*"bill"')

Topic #5 (4, '0.014\*"news" + 0.014\*"one" + 0.014\*"obama" + 0.013\*"supporters" + 0.012\*"watch" + 0.008\*"pence" + 0.008\*"needs" + 0.008\*"golf" + 0.007\*"love" + 0.007\*"right"')

**Pennsylvania:**

**NMF:**

[ 787 1112 752 473 162 918 1074]

people

trumpcare

obama

healthcare

care

right

think

[ 753 1112 752 473 918 1121 1074]

obamacare

trumpcare

obama

healthcare

right

trumptrain

think

[ 627 162 435 57 801 1238 1086]

like

care

going

america

plan

work

today

[1118 918 1074 435 1238 1086 472]

trumprussia

right

think

going

work

today

health

[ 786 752 473 918 57 1086 1171]

pence

obama

healthcare

right

america

today

uniteblue

[ 665 162 435 1054 57 670 1086]

maga

care

going

tax

america

make

today

[ 904 1112 162 1074 1121 435 943]

resist

trumpcare

care

think

trumptrain

going

saveaca

/\*\*//\*\*//\*\*//\*\*//\*\*//\*\*//\*\*//\*\*//\*\*//\*\*/

(8178, 1262)

5 73.47788783198567

10 72.311824761006

15 71.30217390740431

20 70.42016784677571

25 69.62758089075112

**LDA:**

Topic #1 (0, '0.034\*"trumprussia" + 0.024\*"take" + 0.022\*"u2019" + 0.020\*"obama" + 0.017\*"one" + 0.014\*"health" + 0.014\*"could" + 0.014\*"american" + 0.014\*"best" + 0.014\*"fbi"')

Topic #2 (1, '0.020\*"trumprussiacollusion" + 0.019\*"voters" + 0.013\*"man" + 0.013\*"people" + 0.012\*"democrats" + 0.011\*"employer" + 0.011\*"get" + 0.011\*"future" + 0.010\*"another" + 0.010\*"loser"')

Topic #3 (2, '0.029\*"trumpcare" + 0.022\*"obamacare" + 0.020\*"away" + 0.020\*"war" + 0.019\*"bernie" + 0.012\*"something" + 0.012\*"government" + 0.012\*"maybe" + 0.011\*"tweets" + 0.011\*"ahca"')

Topic #4 (3, '0.034\*"via" + 0.022\*"going" + 0.019\*"plan" + 0.018\*"well" + 0.016\*"daily" + 0.015\*"disenchanted" + 0.015\*"aged" + 0.015\*"newsletter" + 0.013\*"would" + 0.012\*"us"')

Topic #5 (4, '0.048\*"like" + 0.030\*"guy" + 0.021\*"message" + 0.013\*"said" + 0.012\*"look" + 0.011\*"golfing" + 0.011\*"spicer" + 0.010\*"human" + 0.010\*"soros" + 0.010\*"worst"')

# Insights

From the sentiment analysis we find that over all there is a positive attitude about the president. However, in states like California the negative sentiments are more predominant while in republican states like Texas the sentiments are more positive and are in favor of the president.

From the word cloud, we find that the topic most tweeted about was health care and the Russian issue all over the US. However, to be more specific Pennsylvania speaks more about Russia rather than health care.

From the topics we obtained from topic modeling the most trending topics are about Healthcare. Despite the geographic location, Health Care is trending because president Trump is planning to abolish the ObamaCare plan and rebuild it as TrumpCare with more additions to it. This discussion has been subjected to massive criticism all over the United States and people have taken to social media to express their views.