A Study of Political Commentaries Online
A Natural Language Processing Experiment

Mark Gallo

A thesis submitted for the degree of
Master of Quantitative Methods of the Social Sciences

May 2020

Abstract

This study aims to explore whether natural language processing can be leveraged to understand the political divide in the United States. Natural language processing (NLP) has long been used to summarize texts and make predictions, through this research I look to understand whether one can use NLP techniques alongside YouTube data to categorize comments by their estimated political ideology, in order to understand the political arena that has emerged from online videos regarding political news. This study looks to test the hypothesis that those with influence in YouTube's comment section will inspire commentaries that support the ideology of the channel the comments are posted to. In order to accomplish this, I will be using the data collected from YouTube's comment section, as no such data is currently available that has the scope necessary to adequately consider the hypothesis. The data collected contains 200,000 rows of data, with variables regarding channel name, user ID, content of comment, like count, reply count, and reply ID. This data will be used to generate measures of a user's ability to generate replies and likes. In this research we describe Social Network measures that we use to understand the ideological divide that may exist on YouTube. It is expected that those who have influence within the comment section will provoke users to make claims in support of the expect ideology of the posting channel. Participation from users of different ideological backgrounds in the comment section of a polarizing channel will then be analyzed to test if users are being guided to communicate with those of alternative perspectives or if the environment is combative and not promoting those to defend their statements as opposed to welcoming opinions of alternative perspectives.

Introduction

The inspiration for this research came from having spent most of my life consuming content from YouTube, as it was free and open to anyone with an internet connection. I believe that access to the internet shaped my life dramatically, but that YouTube played the largest role of all the available platforms, which during the first decade of the 21st century was notably more limited that it is today. During this first decade cites like YouTube, Facebook, and Wikipedia were commonly referred to as dangerous and untrustworthy, Instagram and Twitter as we know them today ceased to exist, and Apple had only just released their first IPhone.

Since then the landscape has changed dramatically. Twitter has become the academic's primary source of quick access social data, YouTube's user count now equates to roughly a quarter of the planet's population, and Apple has long been recognized as the most profitable company in the world, large in part to the success of the IPhone. I was a teenager through this transitional period. I watched as the world changed, I am not at liberty to say that it was for better or worse. Time will untimely make that decision.

Nonetheless I was part of a drastic change, and experiment of sorts. The experiment, as I understand it, appears to ask a fairly simple question. What would happen if we gave as many people as possible affordable access to the opinions of everyone who was willing to share? This research will study what I believe to be the best arena to understand this experiment, and it will focus on what has been understood to be the most divisive portion, American politics.

I refer to YouTube as the best arena because it has most successfully combined the resources that humans use to communicate: the written word, the spoken word, the still image, and the moving image. While platform like Twitter and Facebook grew in prevalence by focusing on the written word YouTube's focus is the moving image. If one were to assume that a picture is worth 1000 words, then one could expect a single five-minute video, shot at 60 frames per second, with an average number of 1800 comments, each roughly 24 words in length, to have an estimated word count of roughly 18,043,200. For comparison the sum total number of words found in the Bible and Quran is less than 1,000,000.

Communicating online has changed how people perceive the world, as it has introduced a plethora of previously unobtainable statements to public. This can be seen through political discourse, which over the past decade has grown in popularity. This research looks to better understand how users are communicating by asking whether natural language processing can reveal political division through the creation of a corpus of words sourced from the commentary itself, as well as political speeches. I hope that in the future we can build off this research, while being conscientious of other peoples' livelihood. Classify a user based on a corpus of words is rather dangerous. It assumes the creator of the corpus can dictate what is and what isn't true of subjective matters. This research will show the successes and failures of such an approach.

Methodology

The literature regarding attempts to understand social media platforms as they relate to political news continues to be updated with time. Some researchers look to understand just the narrative of the comments they are leveraging, while some look to draw comparisons between more traditional data points and those obtained from online forums. Both approaches are useful in their own manner and can bring to light information that exists within the data.

Technologies like Python allow researchers to leverage the Natural Language Tool Kit (NLTK) which includes an extensive supply of functions that one can leverage to conduct text analysis (Bird, Klein, & Loper, 2009). Thanks to such resources it has become simpler to leverage natural language processing techniques to quickly extract information from a data set.

Before attempting to extract information from a data set, one may look to preprocess the data. For text analysis this may consist of tokenizing, removing stop-words, assigning the same case to each text document, stemming, lemmatizing, and or normalizing the text (Kalra & Agrawal, 2018). These preprocessing steps are not required, as some models trained on text data do not benefit from having a reduced sample space (Liu et al., 2019). However, for those who are developing their understanding of a data set, preprocessing is often necessary as fitting a model requires the data to be formatted appropriately.

For example, those who are trying to reveal some characteristics of a document through the words most frequently used, should look to consider tokenizing and removing stop-words. Tokenization reformats the data, and replaces large strings full of uninformative white space and line breaks with lists containing only text and punctuation. Removing stop-words goes one step further by removing words that often appear in texts but add little value. Within NLTK there is a corpus of stop-words that contains a number of these uninformative words. The corpus contains, but is not limited to words like: a, able, above, and almost (Bird et al.). Removing these words can add value to one's summary statistics by making it so that the words which are most critical to the text are included, while those words that add little value are removed.

I elected to remove stop-words for the development of this research, as determining where a comment exists on the political spectrum is unlikely to be made more accurate by the usage of words used in common speech. Ultimately I look to provide a label to each comment and doing so will be based on the usage of the words that have the most relevance within the entirety of the data set.

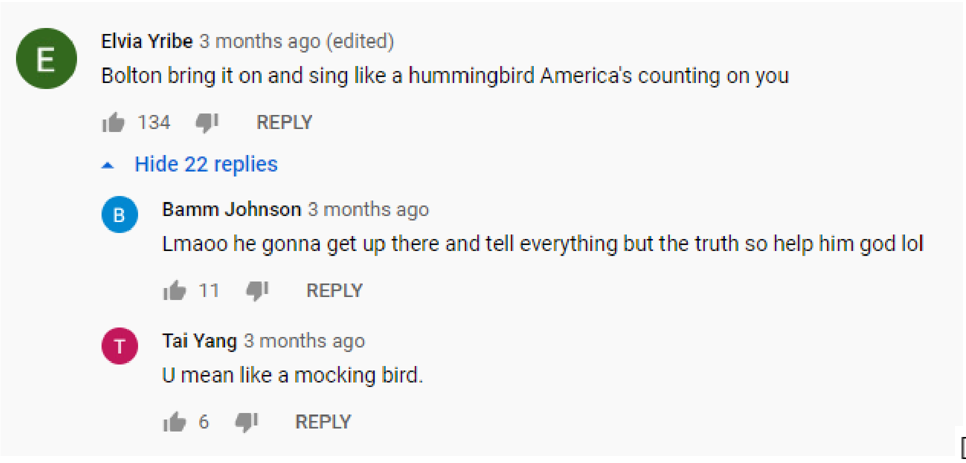
I look to assign labels to comments as labeling the data will make it possible for one to fit a classification model, which will test to see if the content of the comment can be used to accurately predict the assigned label. To perform this labeling we will use Latent Dirichlet Allocation (LDA), "a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. In the context of text modeling, the topic probabilities provide an explicit representation of a document"(Blei, Ng, & Jordan, 2001). I use LDA as an introductory step, so extract words which are meant to represent the conservative and liberal political ideologies in the United States.

After having generated a corpus of words meant to be representative of the conservative and liberal belief systems we will use the word labels to determine if a comment should be labeled as conservative or liberal. A comment that has more liberal words will be labeled liberal, and a comment with more conservative words will labeled conservative. Comments that have an equal quantity of liberal and conservative words will be labeled neutral and words that contain no words labeled liberal or conservative will also be labeled neutral. These labels are what will make the classification of comments possible.

One can perform the extraction automatically through the usage of a text processing pipeline that allows for features to be obtained from documents within the corpus. A feature can be thought of as a label which is, "synonymous of input variable or attribute" (Guyon & Elisseeff, 2008, p. 2). Features are what allow clusters to be created, as they bring to light the similarities that various documents may have, and allow for them to be grouped accordingly. In the context of medical research one would consider a patient's physical symptoms a feature. This is because patients may have a variety of symptoms but this variety allows for clusters to be developed based around the similarities (Guyon & Elisseeff, 2008). Feature extraction is particularly helpful for those using machine learning to study texts, as the features can be used to make predictions regarding the data in question.

Once the data has been adequately prepared, and the desired features extracted, we can focus on the users of interest. As stated previously, the use of social media has grown over the years with large quantities of users signing into platforms like YouTube all across the world at all hours of the day. With this being the case, it is hard to ignore an interest in what users are saying. However, with so many users online, it can be challenging to discern the signal from the noise. In order to assure a clear signal, I focus my interest on those who are most successful at generating responses from other users within YouTube's forums, or as they are more commonly known, the comment sections. These users can be thought of as influencers, but they are not influencers in the sense that readers are accustomed to. These influencers do not necessarily have a large following, brand deals, or get paid to participate on YouTube, but they instead provoke users to respond to them in the comment section. An example taken from a MSNBC video within the data set can be seen in figure 0.

Video: "*Impeachment 'Drug Deal' Witness John Bolton Agrees to Testify at Trump Trial, Defying WH | MSNBC*"



Elvia Yribe 3 months ago (edited)
Bolton bring it on and sing like a hummingbird America's counting on you
134 REPLY

Hide 22 replies

Bamm Johnson 3 months ago
Lmaoo he gonna get up there and tell everything but the truth so help him god lol
11 REPLY

Tai Yang 3 months ago
U mean like a mocking bird.
6 REPLY

[0]

Upon observing the channel statistics of the original commenter one can deduce that this account is not associated with what one would typically consider to be an influencer. However, within the comment section we can see that this user has inspired 22 replies and 134 likes. Upon further investigation this comment appears to be well supported by those within the forum.

In order to realize who, the most influential users in a forum are, I observed their performance statistics. This approach was inspired by the way sports analysts study players, as ultimately one wants to see an influencer generate responses, like how an athlete generates points and victories. I focus on the number of comments inspired, the number of likes received, and the number of comments an influencer posted to a forum. I observe the number of comments posted, as a measure of how active a user is within the forum.

Once the influencers have been sorted and their inspired commentaries properly labeled, I look to see if the most influential users are inspiring conversations regarding key issues, are creating ideological bubbles, or are effectively shouting into a void. Performing this analysis allows me to address the key research question of this report. If it is found to be possible that one can determine, using unsupervised learning techniques, that the most influential users are generating ideological bubbles than this research would support the belief that those who participate in American politics, and express their stances in YouTube comment sections, are either in support of or against the ideological beliefs of the channel. If diversity of ideology is present, meaning an influencer generates comments that are labeled both conservative and liberal than this may have found support for the belief that those who participate on political forums are willing to discuss both sides of an issue, as they are willing to be a part of social networks that contain users of different ideological perspectives.

Theoretical Framework

As internet usage has grown, so has the usage of social media platforms. Such platforms have found their way into the political realm, as users have sought to leverage social media to get closer to candidates, allies, and those with differing viewpoints.

Social media sites, like Twitter and YouTube, have become the safe havens for those who wish to express their political viewpoint publically (Bollen, Pepe, & Mao, 2009). With so much information being shared by users, social media sites have become public digital forums, where one can observe, consider, and discuss almost anything. This change can be partially attributed to the permanently public nature that comments online have. An example of this characteristic can be found in the backlash noteworthy individuals receive when a post is deemed insensitive by a community of users. The backlash may not happen as soon as the post is made, as it may only occur once public opinion has shifted regarding an issue (Andrews, 2018). This is possible because posts on social media platforms do not disappear naturally with time, for a post to no longer be found online it must be removed, either by the user or a mediator. The rules regarding posting etiquette vary depending on what cite one is on. Some cites provide documentation so users can know what is appropriate, but they tend to be for content creators or those who produce services that can be advertised on. Comment forums will sometimes assign a moderator(s) to oversee the content posted on behalf of the cite or forum. The moderators job is to make sure the content exists within the created guidelines (Reddit, 2017). YouTube has documentation regarding what its creators are allowed to post in order to assure monetization, however this documentation does not discuss the comment section of a creator's videos. (Google, 2020).

Previous research has found that YouTube comments are assumed to be, "inadequate", "juvenile", or containing text meant to inspire hate (Schultes, Dorner, & Lehner, 2013) . Despite these harsh critiques, YouTube's audience has continued to grow, as of 2020 YouTube had more than 2 billion users, in over 100 countries, who together speak 80 languages, and watch more than one billion hours of content daily (YouTube, 2020). These figures are quite impressive, especially when one considers than in 2013 YouTube had only "800 million users" (Schultes et al.). YouTube was able to increase its already tremendous user base by 150% in less than a decade, making it so that roughly one third of the internet can be classified as having a YouTube account (YouTube). With so many people participating on the platform over the past decade the cite has become a source of a tremendous amount of anecdotal information.

Past research has found that it is possible to extract an emotion from comments found on YouTube using either Support Vector Machines or Pointwise Mutual Information parameters. Findings have shown that both approaches perform well, with accuracies above 68% and precision falling just under 93% (Douiji, Hajar, & Hassan, 2016, p. 297). Pointwise Mutual Information requires that a set of categories be made, each designed to represent a potential human emotion. A set of words that represent the given emotion will then be used to populate documents that the model can use to train. One could use the emotion of happiness as an example. Happiness can be represented through a number of words, some particular ones are great, fantastic, enjoyable, and content. What one now has is a list of words that are symbolic of the emotion happiness. This allows one to measure how prevalent a characteristic like happiness may be in a given text, by using the words within a corpus as signifiers. If a word were to not appear as an emotional signifier, but does an above average job of expressing a feeling close to the emotion of interest, for example jovial was not on the list provided above, but linguistically does a reasonable job of expressing an emotion that resembles happiness, it should be added to the list of emotional signifiers. Assuring that the set of signifiers is updated allows for adjustments to be made that cater to the way people are using language to create text. As the nature of a commentary changes it is important that the corpus used to analyze such commentaries is adjusted to suit the ever changing idiosyncrasies of social interaction.

The YouTube population has experienced tremendous growth over the last decade. Initially one may expect to realize increased quantities of data, and better findings as a result. While this may be the case, it is important to note the amount of potential noise that exists within the data set. Previous research has found that roughly 30% of comments can be labeled as an anecdote that correlates to the topic being discussed or referenced in the video (Schultes et al.). This leaves 70% to be classified as noise, or data points that do not guide the researcher to a useful conclusion. The data preprocessing stage of this research is a necessary step as I look to limit the potential noisiness of the data. This is the stage where the sample data is studied and documents with low value are removed, so that the study can be performed on a robust and informative data set.

There are a number of important questions that one can ask regarding the usefulness of YouTube comments. It is important that through this research an understanding of the data is developed, as to allow for the construction of a function that can adequately explain some of what makes a text useful. Some previously asked questions are: “Can we predict the community feedback for comments? Is there a connection between sentiment and comment ratings? Can comment ratings be an indicator for polarizing content? Do comment ratings and sentiment depend on the topic of the discussed content?” (Siersdorfer, Chelaru, Nejdl, & Pedro, 2010). This paper will look to address these questions as they relate to the political discussion that has been accumulating within the comment section of political videos that exist within conservative and liberal minded playlists.

Once one has preprocessed the data and has it ready to be analyzed there are a number of traditionally leveraged techniques that one can apply, all of which are useful in their own manner. The first task one will often look to take on is document classification. This makes the creation of clusters that allow for useful visualizations possible. Clusters are relied upon because a single data point may not be able to inform a researcher of how the masses are reacting to a given occurrence, but many data points of similar nature clustered together by similar characteristics can prove useful, as it allows for a quantitative understanding to be developed.

As previously stated, YouTube data is often assumed to be very noisy, and full of extra jargon and short-hand that makes drawing conclusions through text analysis challenging. This challenge is associated with the need to make so many adjustments to the underlying data.

Once it is understood that one can leverage data from social media, use it to conduct analysis, and determine that the analysis has been performed appropriately, one is left to ask why one would go through all the trouble. One may be able to generate information about those who use YouTube to watch videos on the discussed topic, but has gathering this information during past research attempts led to worthwhile findings?

Past research has found that content online can influence one's beliefs, but it depends how the person to be influenced participates on the platform. This can be interpreted in one of two manners. The first interpretation is dynamic, which suggests that those who participate online and closely identify with a particular comment will show how they connect to the anecdote through increased online participation. (Walther, DeAndrea, Kim, & Anthony, 2010). This can be observed through revealing that a comment left on a video received a notable amount of feedback from a user and this feedback is associated with a certain sentiment. This approach suggests that those who participate online are dynamic creatures whose perceptions can be altered by the statements they associate with online.

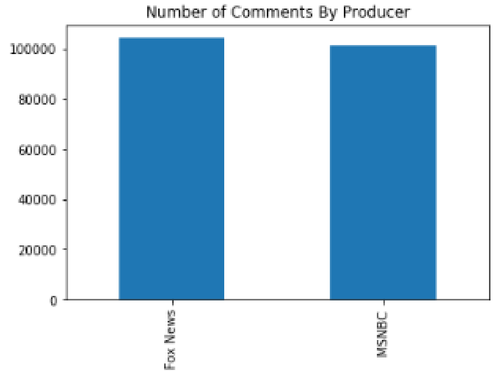
The static interpretation assumes that participants are not so easily molded. One who subscribes to the static interpretation would suggest that regardless of a given stance, one will have a reaction to the content provided in the video, and then seek out content that align with the individual's beliefs on the issues. If one were to strongly agree with a message that exists within the observed content, then he would likely seek out comments that support his belief. The static interpretation assumes people will stay true to character and that the influence of content is not strong enough to shift a person's stance on a given issue, thus it is more frequently used to add support to ideas individuals already believe (Walther et al., 2010).

This research will focus on the most influential users as well as those whom they influence. This is important to note because the dynamic and static interpretations of how users perceive online content will be observed through the leveraging of a ranking system. While previous research seems to suggest that those who align with the same political opinion will agree with those of the same mindset, while agreeing to disagree with those of a different one, it is important to understand what the issues are that are being presented and whether or not users of diverse ideological backgrounds are actively discussing them (Barberá, 2015). This research looks to determine if ego-chambers exist and if so, who or what is creating them? Could a platform that has a massive user base, that is growing seemingly every day, begin to cater itself toward more useful and productive dialog? Has the increase in user base inspired users to be more respectful, in such that it has made them more considerate about what should be said online? Questions of this nature as well as the one's previously discussed will be addressed through this research, where ultimately the goal is to understand the ideological ego-chambers that have developed as a result of content being delivered in a manner that allows for public discourse to develop in a segregated space.

This research intends to leverage models that will better inform us as to what is being discussed in these forums, the nature of the discussion being had, the number of individuals participating in the discussion, and measures of how impactful a given anecdote was for the collected sample. This will be done through the rating system that all comments on YouTube are subject to once posted. It is a mostly democratic system as those comments with the largest number of supporting votes tend to exist near the top of the forum. One should note that time is also an element of the function that determines where in the forum a comment resides. This means that comments that garner a substantial amount of support will be near the top of the forum so long as time does not permit another comment to climb the ranks. Some comments that have been posted for longer than others will be moved down to introduce new comments to the forum, which may attract support. The manner in which YouTube displays comments has been considered in this research, as the collection approach required that all comments be taken from a video's forum, so as to assure that all candidates for most influential were being considered.

Data and Variables

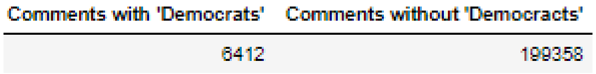
Before referring to the results of the models I will discuss some simple summary statistics regarding the data set, so to have an understanding for the analysis that will be performed.



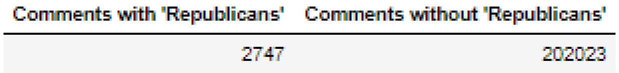
[1]

The first visualization is a bar plot to show the size of each grouping of comments. The comments are split by channel name. This project refers to the popular political commentary channels Fox News and MSNBC. It can be seen by observing figure 1, that there is a similar quantity of comments sourced from Fox News and MSNBC. One may be inclined to operate under the assumption that comments left on Fox News videos could be labeled conservative, and those posted on MSNBC videos could be labeled liberal. This research does not leverage that assumption, as the focus is on the commentary occurring within the comments. To operate under this assumption one would need to assume that all statements which occurred as a result of MSNBC would be liberal, but previous research suggests that this is unlikely to be the case. When studying Twitter in a similar manner it was revealed that most Twitter feeds are full of critical commentary regarding the political figure of interest. This translates to a figure like the United States' 44th president Barack Obama having a Twitter feed which consists primarily of comments that lack support for President Obama, the same can be said of figures on the right side of the political spectrum (Barberá, 2015).

To begin my analysis, the Python programing language was used to determine the presence of key words and phrases, in order to see how often they appear in the data set. The first key words of interest are the names of the two major parties in American politics.



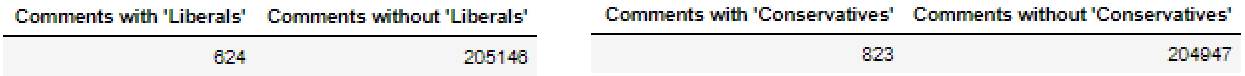
Comments with 'Democrats'	Comments without 'Democracts'
6412	199358



Comments with 'Republicans'	Comments without 'Republicans'
2747	202023

[2]

One can see from figure 2 that the word *democrats* appears in 6,412 of the comments, equating to roughly 3%, and *republicans* appears in 2,747 equating to roughly 1%. With regard to the words *liberals* and *conservatives* it was found, and shown in figure 3, that both words appear in less than 1% of comments.



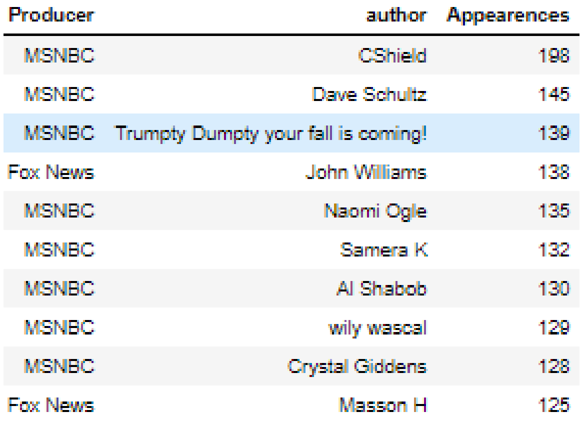
Comments with 'Liberals'	Comments without 'Liberals'
624	205146

Comments with 'Conservatives'	Comments without 'Conservatives'
823	204947

[3]

Because it was found that some of the commonly used words when discussing politics are not frequently found within the data set a technique that considers the nature of the comments themselves will need to be employed, so to best assure that each comment is being assigned the appropriate label.

While the results of this research will not be based on the assumption that the location of a comment can have a measurable impact on where the comment is positioned on the political spectrum, it will leverage the location of the comment posting as a tool to help develop an understanding of the data.

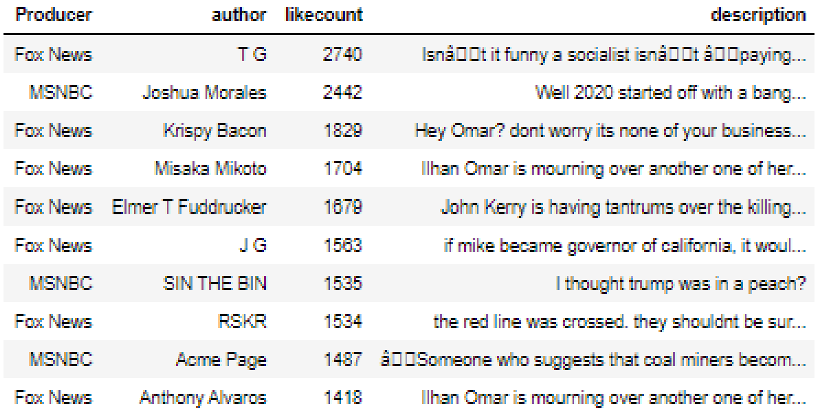


Producer	author	Appearences
MSNBC	CShield	198
MSNBC	Dave Schultz	145
MSNBC	Trumpty Dumpty your fall is coming!	139
Fox News	John Williams	138
MSNBC	Naomi Ogle	135
MSNBC	Samera K	132
MSNBC	Al Shabob	130
MSNBC	wily wascal	129
MSNBC	Crystal Giddens	128
Fox News	Masson H	125

[4]

Figure 4 considers which commenters post the most comments to forums regardless of which channel the comments were posted to. It can be seen from those who appear in the top ten of appearances that most of the commenters are found posting on MSNBC videos, with 80% of them choosing to post on those particular forums.

Figure 5 is used to understand the data set by showing those users that were able to generate the most likes from a single comment. From this listing we get a significantly different perspective of the data. While MSNBC commenters dominate the posting sphere, the ability to generate likes is led by posters to Fox News videos. It can be seen that the most liked comment was posted to a Fox News video and generated 2,740 likes. One can observe that in the top ten, 70% of the top like generating comments were from the Fox News YouTube channel.



Producer	author	likecount	description
Fox News	T G	2740	Isnâ□□t it funny a socialist isnâ□□t â□□paying...
MSNBC	Joshua Morales	2442	Well 2020 started off with a bang...
Fox News	Krispy Bacon	1829	Hey Omar? dont worry its none of your business...
Fox News	Misaka Mikoto	1704	Ilhan Omar is mourning over another one of her...
Fox News	Elmer T Fuddrucker	1679	John Kerry is having tantrums over the killing...
Fox News	J G	1563	if mike became governor of california, it woul...
MSNBC	SIN THE BIN	1535	I thought trump was in a peach?
Fox News	RSKR	1534	the red line was crossed. they shouldnt be sur...
MSNBC	Acme Page	1487	â□□Someone who suggests that coal miners becom...
Fox News	Anthony Alvaros	1418	Ilhan Omar is mourning over another one of her...

[5]

So far it can be understood that, given this data set, those comments that are able to generate the most likes tend to be from users who comment on Fox News videos, while those who comment on MSNBC appear to be doing so more frequently. While it is interesting to know who is leaving the most comments and which of those comments are getting the most likes, for the purpose of this research it is critical to know which users are generating responses from others in the data set.

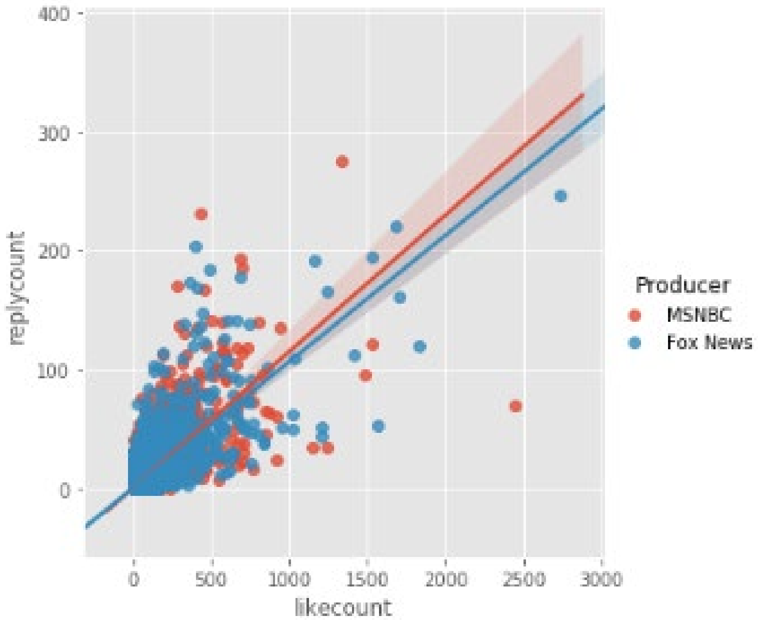
Observing figure 6, which can be seen on the following page, one can see that the ability to generate responses, in terms of where the comment was posted, shows the greatest amount of diversity. It can be seen that with regard to generating replies 60% of the top users posted to Fox while 40% posted to MSNBC. When looking into the nature of the comment left, one quickly realizes that the comments that generated the most replies were attacks on the party that would not typically be associated with the channel the comment is being posted to. For example, the top comment, which generated 276 replies was, "Republicans can only create two things: deficits and war." This was posted to the MSNBC comment forum. The second most replied comment was left on the Fox News forum and read, "Isn't it funny a socialist isn't paying her fair share?". These top comments seem to suggest that what was found in reference to Twitter data (Barberá, 2015) may be generalizable beyond the Twitter ecosystem. What is found in reference to YouTube comments is that those generating the most responses tend to be "attacks" toward the side of the political spectrum that the channel producer would be unlikely to support. Suggesting that on YouTube those comments which side with the producer are more likely to be noticed, a contrast from what was seen during the previous analysis of Twitter. What is similar between these two analyses is that the comment sections, at a surface level, are safe spaces for those who wish to slander their political opponents. Going further this paper will consider how safe these spaces are, as spaces that lack ideological diversity would be likely to be safer places to post distaste for alternative viewpoints.

So far it can be understood that, given this data set, an individual Fox News comment is likely to be able to generate the most likes, while an MSNBC comment was able to generate the most responses. However, Fox News commenters are able to generate a nearly equal response, if one aggregates the success users have had generating responses. This research operates under the belief that those who are capable of generating responses have an inherent value, that can shape the nature of the comment forum and drive the commentary in a given direction. Because

YouTube is such a large platform, where individuals, and organizations can post a wide array of content, it is important to understand that not all comment sections will have near similar content characteristics. Thus not all comment sections should be expected to develop in a similar manner. However, when disregarding typical measures like views and subscriber counts, so to compare two American political channels based on their similarity of genre, it has been revealed that the number of likes, comments, replies, and type of comment posted are quite similar. This is a finding that will be discussed in greater detail throughout the body of this research.

Measures / Results

Likes are often referenced as a measure of the popularity of a post on social media, while replies are not discussed as frequently. Figure 6 reveals, regardless of channel, likes and replies trend in a similar direction. This trend appears to produce a Pareto distribution, as most comments generate a small number of responses and do not receive many likes, while a select minority can generate a disproportionately large quantity of each. It can now be seen that the channels have similar trends with regard to comments, as the lines of best fit appear to have

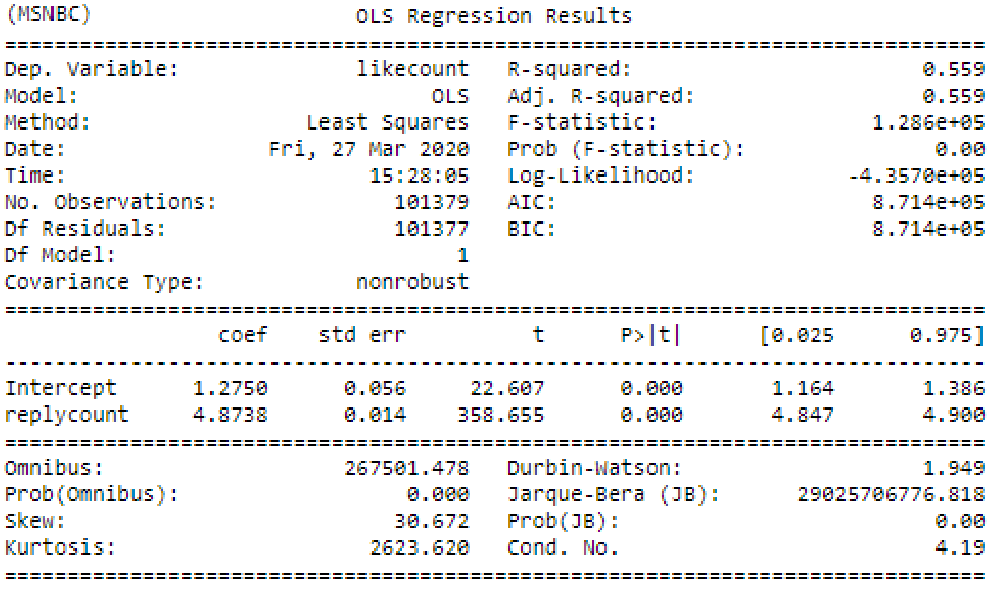


[6]

similar slopes. Note that this is true despite Fox News having 2 million more subscribers than MSNBC.

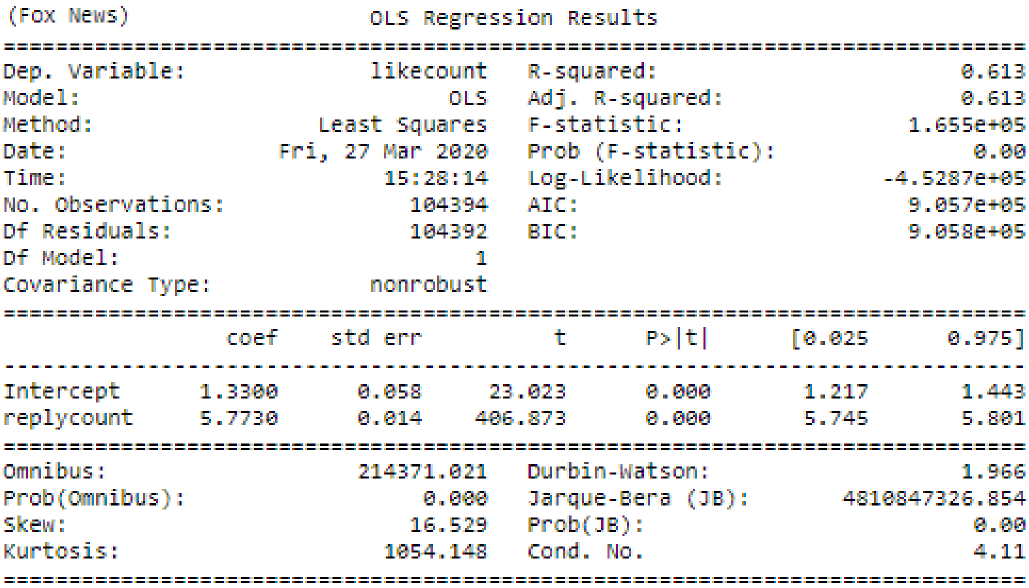
After seeing this one may be inclined to question what can be explained by a channels subscriber count. Commonly a subscriber count is referenced as a representation of the number of individual accounts that support a channel. However, this data set appears to provide an example that allows one to see that when generating engagement, subscriber count may not provide much explanatory power, as having 100% more subscribers did not result in an average change in the number of replies a comment received or likes it got. This could be explained in any number of ways; however, the prevalence of bots or inactive accounts could partially explain how it is possible to have such a substantially larger subscriber base without realizing significantly more activity from users.

The regression output for MSNBC comments can be seen in figure 7, it is realized that when commenting on an MSNBC video, one can expect to receive, on average, 4.9 likes for each additional comment reply. A t-score of 358.65 was realized suggesting that one can reject the



```
(MSNBC)                     OLS Regression Results
==============================================================================
Dep. Variable:              likecount   R-squared:                       0.559
Model:                            OLS   Adj. R-squared:                  0.559
Method:                 Least Squares   F-statistic:                 1.286e+05
Date:                Fri, 27 Mar 2020   Prob (F-statistic):               0.00
Time:                        15:28:05   Log-Likelihood:            -4.3570e+05
No. Observations:              101379   AIC:                         8.714e+05
Df Residuals:                  101377   BIC:                         8.714e+05
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      1.2750      0.056     22.607      0.000       1.164       1.386
replycount     4.8738      0.014    358.655      0.000       4.847       4.900
==============================================================================
Omnibus:                   267501.478   Durbin-Watson:                   1.949
Prob(Omnibus):                  0.000   Jarque-Bera (JB):      29025706776.818
Skew:                          30.672   Prob(JB):                         0.00
Kurtosis:                    2623.620   Cond. No.                         4.19
==============================================================================
```

[7]

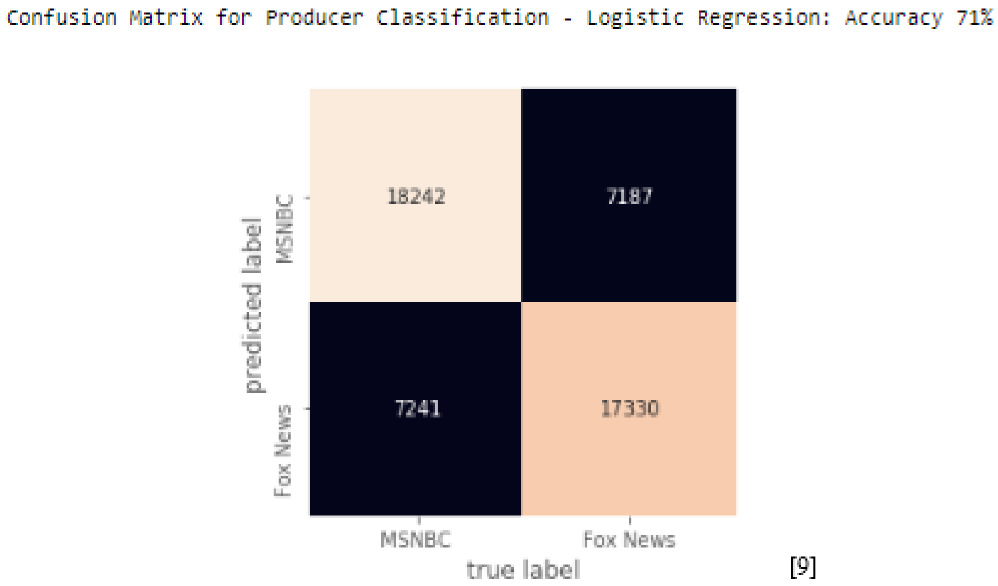


```
(Fox News)                 OLS Regression Results
==============================================================================
Dep. Variable:              likecount   R-squared:                       0.613
Model:                            OLS   Adj. R-squared:                  0.613
Method:                 Least Squares   F-statistic:                 1.655e+05
Date:                Fri, 27 Mar 2020   Prob (F-statistic):               0.00
Time:                        15:28:14   Log-Likelihood:            -4.5287e+05
No. Observations:              104394   AIC:                         9.057e+05
Df Residuals:                  104392   BIC:                         9.058e+05
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      1.3300      0.058     23.023      0.000       1.217       1.443
replycount     5.7730      0.014    406.873      0.000       5.745       5.801
==============================================================================
Omnibus:                   214371.021   Durbin-Watson:                   1.966
Prob(Omnibus):                  0.000   Jarque-Bera (JB):       4810847326.854
Skew:                          16.529   Prob(JB):                         0.00
Kurtosis:                    1054.148   Cond. No.                         4.11
==============================================================================
```

[8]

null hypothesis that the number of replies a comment receives has no influence on the number of likes it will get and the 99% confidence interval. An adj. R-squared of 0.56 suggests that once the size of the data set is accounted for, 56% of the variation in the number of likes a comment received can be explained by the number of replies it generated. Fox News data produces near identical results, which can be seen on figure 8. One can see that the statistical outputs produce near similar coefficients, adj-R squared values, and t-scores, suggesting that this trend occurs regardless of channel. I find this particularly interesting as it suggests that if one were to assume that each channel is representative of an ideological framework then the two sides, which seldom align on beliefs and values, would be found to behave quite similarly to one another in terms of engagement with content.

To further understand the data, I chose to see whether common machine learning techniques could be leveraged to predict which channel a comment came from. This is being done to reveal the differences between the data, as a model that would be able to successfully classify comments according to their posting location would suggest that the comments left on MSNBC videos have characteristic differences from those left on Fox News's channel. To perform this analysis a random sample of 50,000 comments was put through a pipeline so to apply a Tfidf Vectorizer to the text, and classify using logistic regression.



[9]

After applying GridSearchCV to uncover which variable parameters would produce the model with the greatest predictive power an accuracy score of 0.69 was received. The confusion matrix for this analysis can be seen as figure 9. The model performs similar to past studies which have used classification to understand online political commentary (Douiji, Hajar, & Hassan, 2016), and appears to suggest that a majority of comments can be easily classified as belonging to MSNBC or Fox News. However, a sizable proportion of comments are not so easily classified. This may suggest the presence of comments on one channel that would typically be believed to exist on the other. If this is the case then one could begin to understand that the content foun [9] within the comment sections of ideologically segregated YouTube videos may not be as different as one is led to believe, as the incorrect categorizations would suggest that within each forum there are a number of comments that do not fit the mold of the channel they are posted to.

In order to determine whether this ideological segregation is occurring one would need to study the content of the comments themselves, beyond just whether or not the comment would typically exist on a given channel. In order to accomplish this, I employ Latent Dirichlet Allocation to create groupings of words that exist within the comments. These groupings will be based on the subjective, ideological frameworks associated with American politics. Those frameworks are: *Liberal* and *Conservative.* I will leverage a corpus method to accomplish this. After a list of words has been determined the words will be used to label the words within each comment in the data set. These labels will be used to determine whether a comment should be labeled *Conservative* or *Liberal.* A *Neutral* label will be assigned to those comments that are not determined to be within one of the two frameworks of interest.

This research focuses on the most influential actors within the data set, so the labeling will be done specifically on the comment replies that were inspired by the most popular comments. The comments that receive the greatest number of replies are those that generate the most engagement, and require the replying user to perform the most work. It was determined that replies would be considered in place of likes because of the additional effort and thought that is necessary to post an original comment to a forum. It is believed that these users commit a greater sum of time to being engaged with the content and are thus the more influential and influenced users in the forum.

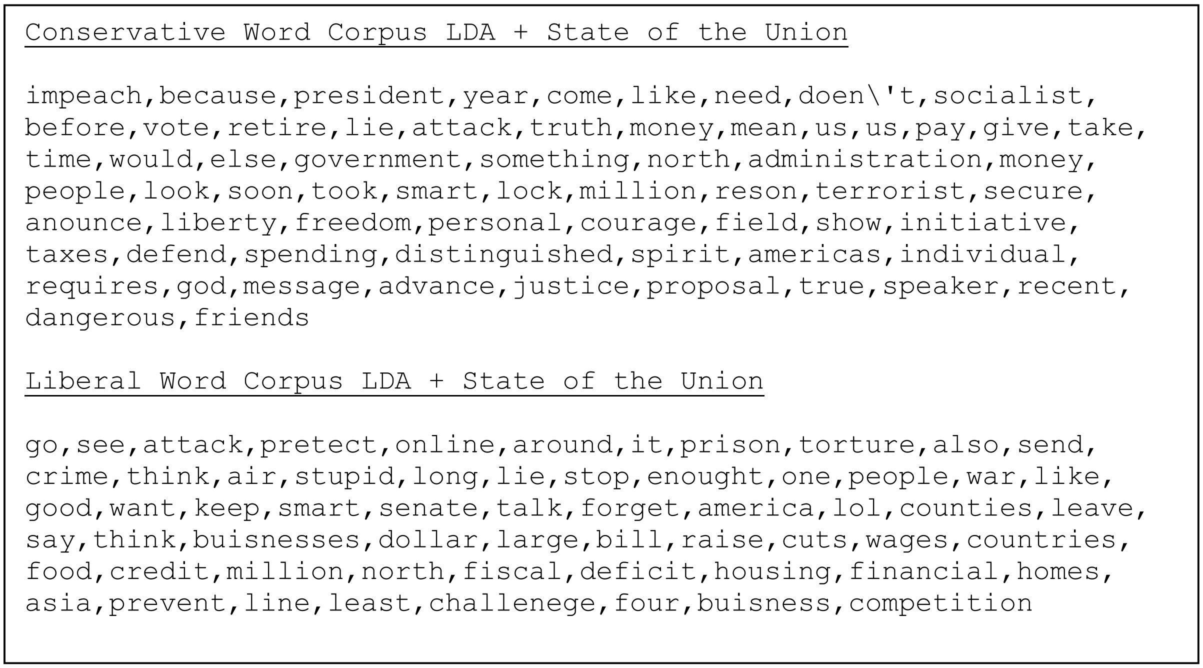
To address the key research question of this project those who are more likely to be involved, influential, or influenced will be studied. The goal is to see if these users are commenting in highly segregated spaces. A space would be deemed highly segregated if it consists of only one ideological leaning. An example would be a comment that receives 10 replies all of which receive the label of *Liberal*. In the provided example only liberal replies would be added to the forum, thus creating a space where only liberal ideologies are being expressed. If this is the case then this paper would provide additional support for the hypothesis that, those who comment on politically charged content likely do so when surrounded by comments of a similar nature. Suggesting that the commentaries being inspired from political YouTube content are ideologically segregated. If it were true that the 10 replies consisted of some liberal, some conservative, and some neutral, then this example would provide support for suggesting that such commentaries are not ideologically segregated.

This research employs a digital labeling strategy to determine whether a comment is *Conservative* or *Liberal*. In order to accomplish this labeling two lexicons were created. One for conservative words and one for liberal words. The creation of these lexicons leverages two approaches, the first being latent dirichlet allocation. This allows the lexicons to contain words that are representative of concepts that are present within the data. This approach generated 20 words for each lexicon, each of which has been classified as *Conservative* or *Liberal.* To determine whether a set was liberal or conservative a uClassify “Liberal or Conservative” classifier was used (Politimind). The output determined that the first set of words was *Liberal*, while the second set was C*onservative*. Upon reading through the outputted words, it was determined that additional words would need to be added in order to effectively assign the labeling of conservative or liberal to a given comment. The words generated from the LDA

```
Liberal Word Corpus LDA
go,see,attack,pretect,online,around,it,
prison,torture,also,send,crime,think,air,
stupid,long,lie,stop,enought,one,people,
war,like,good,want,yet,keep,smart,senate,
talk,forget,america,lol,counties,leave,say


Conservative Word Corpus LDA
impeach,because,president,year,come,like,
need,doen\'t,socialist,before,vote,retire,
lie,attack,truth,money,mean,us,us,pay,give,
take,time,would,else,government,something,
north,administration,money,people,look,
soon,took,smart,lock,million,reson,
terrorist,secure,anounce
```

modeling can be seen below, they have been split into their respective groups, which can be seen in figure 10.

In order to assure that each lexicon captures a *Conservative* or *Liberal* sentiment an additional corpus needed to be leveraged. One that had words that could be directly tied back to a conservative or liberal figured head. To accomplish a lexicon of words that were commonly used in 71 State of the Union Addresses by Democrats and Republicans was leveraged (Rob, 2014). This adds additional credibility to labeling device, as it does not require the comments to be classified from only the content of the comments themselves. The additional words can be found in figure 11.



Conservative Word Corpus LDA + State of the Union

impeach,because,president,year,come,like,need,doen\'t,socialist,
before,vote,retire,lie,attack,truth,money,mean,us,us,pay,give,take,
time,would,else,government,something,north,administration,money,
people,look,soon,took,smart,lock,million,reson,terrorist,secure,
anounce,liberty,freedom,personal,courage,field,show,initiative,
taxes,defend,spending,distinguished,spirit,americas,individual,
requires,god,message,advance,justice,proposal,true,speaker,recent,
dangerous,friends

Liberal Word Corpus LDA + State of the Union

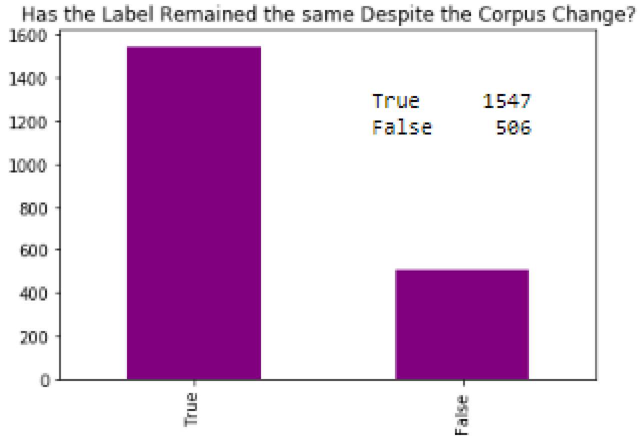
go,see,attack,pretect,online,around,it,prison,torture,also,send,
crime,think,air,stupid,long,lie,stop,enought,one,people,war,like,
good,want,keep,smart,senate,talk,forget,america,lol,counties,leave,
say,think,buisnesses,dollar,large,bill,raise,cuts,wages,countries,
food,credit,million,north,fiscal,deficit,housing,financial,homes,
asia,prevent,line,least,challenege,four,buisness,competition

[11]

After having completed the building of the lexicons the labeling approach requires labeling each word within a comment according to the lexicon. To do this the first step was tokenizing the comments so that each was broken up into its individual words. After the comments were tokenized a loop was used so to apply a label to each word in each comment. The resulting data set contains the comments used, the source the comments came from, and the predicted ideological label. This additional variable can now be used to determine how segregated a grouping of comments is.

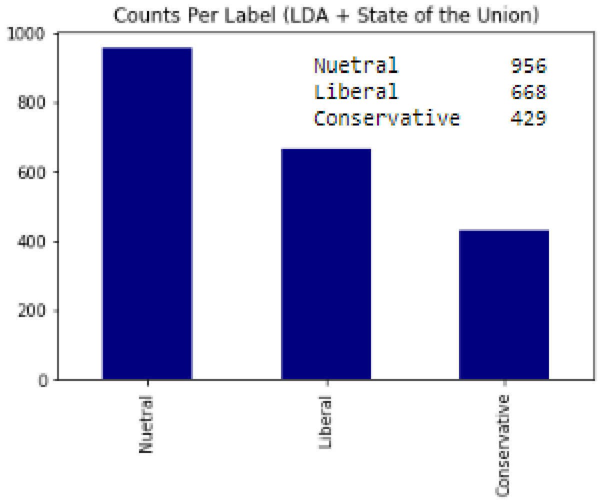
Having introduced the new set of words into the corpus creates change within the labeling results. We can see from the graphic below that of the 2053 total comments 506 of them experience a label change from the introduction of the additional corpus, this equates to roughly ¼ of the comments receiving a new label as a result of introducing additional words.

Having realized that a large proportion of the comments experienced a label change inspired a further investigation into the labeling mechanism. One can see from figure 13 that the introduction of additional words into the corpus inspired and increase in the number of comments labeled *Conservative* and *Liberal*. This makes sense to a degree, as one could understand that having an increased number of words in the lexicon provides a greater opportunity for a comment to receive a label, as there are now more words that can be matches. It is possible that an increase in words could inspire more neutral ratings as the number of ties could increase with the size of the corpus, however given this data that seemed not to be the case.



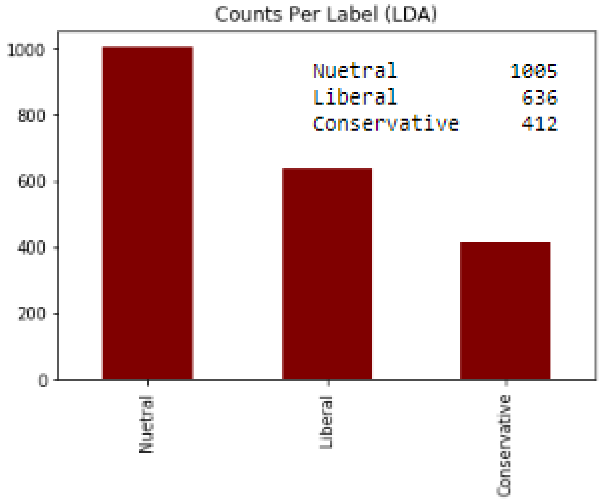
[12]

One realization was that the percentage increase in the number of *Conservative* and *Liberal* labels was small when compared to the number of changes the additional words inspired. The corpus size increase inspired roughly 25% of the words to change labels, but the number of labels being conservative or liberal increased by only 2.5%. This suggests that there is significant variation between the two labeling techniques despite each of them producing a similar number of *Conservative* or *Liberal* labels.



[13]

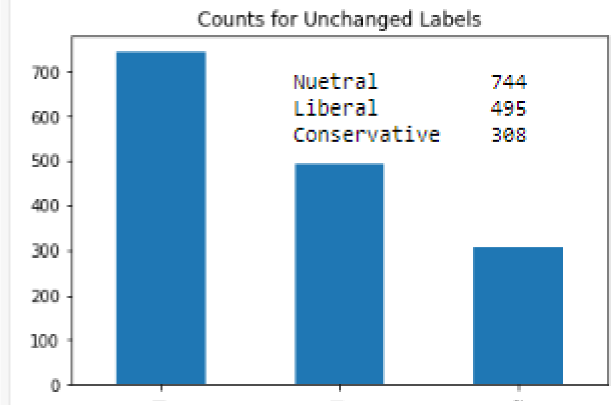
This variation inspired further investigation. From this point I decided to dig deeper into the nature of the comments that did not change with the increase in lexicon size and those that did.



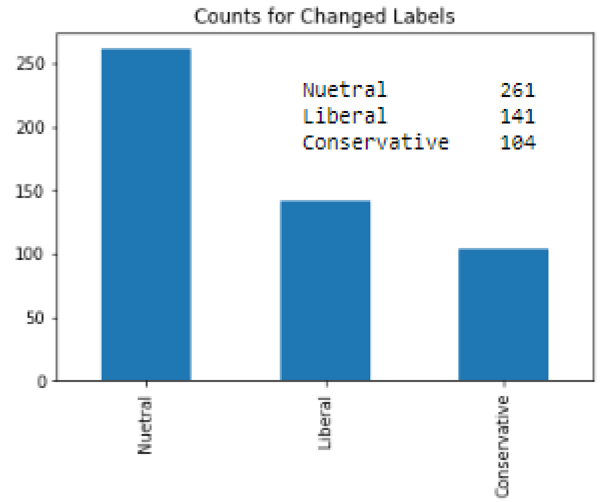
[14]

When looking at those data points that did not change, which can be seen in figure 12, we see that of those points where the label remained the same most of the comments (48%) are neutral. The remaining 52% are either conservative of liberal, with most of those tending toward the liberal ideology.

The graph which can be seen in figure 15 looks specifically at the comments that realized a change in labeling. The following, figure 16, looks at those comments that remained the same. For both a similar trend emerged. Most of the comments are neutral and of the liberal and conservative comments, and the ones that are *Liberal* outnumber those that are *Conservative*. There continues to be an increase presence of *Liberal* labeling. This trend emerges despite the conservative corpus containing five more words than the liberal. This appears to support theories



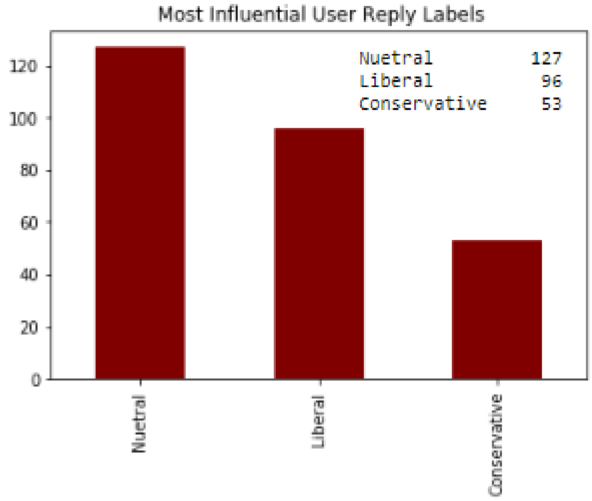
[15]



[16]

that those of *Liberal* perspective are more likely to be active in online political arenas than their *Conservative* counterparts.

Having discussed the data in general it can now be looked at in relation to the presence of a particular ideology. Within the appendix the remaining top 10 most influential users can be found. The most influential has been provided as an example below. Upon consulting the output one can see that the comment inspired over 250 replies, was posted to an MSNBC video, and was labeled *Liberal*.



[17]

The content of the comment was, “Republicans can only create two things: deficits and war”. Upon reading this, it appears that the comment would be associated with the *Liberal* political ideology, as the comment directly calls out Republicans for being bad spenders and supporters of war. Someone of conservative ideology would be unlikely to make such claims of Republicans because Republicans are typically associated with the conservative mindset. It is also noted that the poster of the original comment did not actively participate with those who were replying to the original comment. This suggests that this single comment was able to activate a number of users within the arena and inspire them to begin making claims within the forum.

I then create a visual describing the labeling of the replies that were inspired, it can be seen as figure 17. It is seen that, as expected most replies are neutral, and most of the rest are *Liberal*, this makes sense as the channel the comment was posted to has a *Liberal* tendency and the comment that inspired the replies has *Liberal* characteristics. By looking at some reply examples, it is possible to reveal that the labeling mechanism performs fairly well, while not perfect. The following comments were interpreted as correctly labeled.

Liberal: *It can be blamed on the country as a whole. Starting with the Government, Parents, Poverty, and the NRA.*

Liberal: *If we went to war trump would be cooked.*

C**onservative:** *No it is not! He was a General not a Diplomat. Of course Iran will retaliate. The point what if the U.S. did nothing and was able to a U.S. Diplomat, you would probably apologize to Iran for our Diplomat being in the Middle East to begin*

Conservative: *All of you are in imagination land, actually listening to the opposition of our country through the ELITE owned news.*

However, because no corpus is perfect, some of the comments are mislabeled. I have provided some examples:

Liberal: *LEFTIST ARE REALLY DUMB U DID NOT WANT THAT TO BE A JOKE. INSULTING IGNORANT ASSWHIPE*

Conservative: *I beg to differ. Both Democrats and the GOP a.k.a. Greedy Old Phuckers have engaged in "deficits and wars!" The U S system of governance is DEFUNCT, DEPRAVED,* *DISGUSTING, DEGENERATE, ARROGANT, SHAMELESS, DECADENT AND DYING! Eat your hearts out!*

It is difficult to compute the margin of error on these labels, as they are subjective, but so far it has been found that the labeling mechanism performs relatively well, of the 10 most replied to comments, 9 were estimated to be accurate to the label. This is of course subjective, but given the polarized nature of American politics it is believed that the estimations are within reason. This conclusion was drawn using some key assumptions.

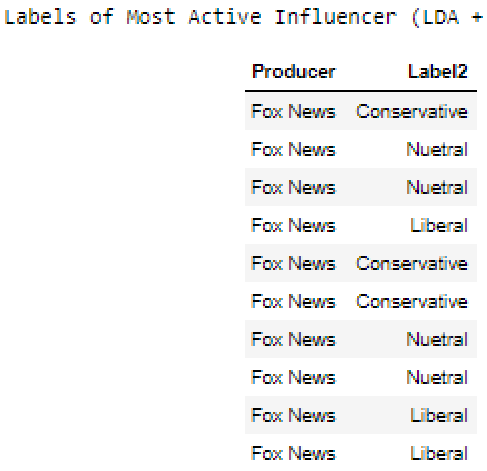
It is first assumed that the poster was sincere, sarcasm is an obstacle of this research, and it was mitigated by reducing the sample size and reading over the comments to check whether the labeling mechanism was missing some context that a human could easily have picked up.

The second assumption is that where a comment is posted can explain some of the content, and potentially reveal some present sarcasm or context. This is why the classification was performed with regard to producers. In doing so it was realized that some unique traits could be found within MSNBC comment forums that would not be present to those of Fox News, and vice versa.

Third is that this research assumes the labels to be accurate, however it is necessary to proceed with an understanding that no corpus can perfectly label an ideology, as perceived ideology is highly subjective. If one were to consult comments from political videos of a different country the label of *Conservative* and *Liberal* may not make much sense given the context of the political arena.

Going through the appendix one can observe the response labeling of the most inspirational comments and the replies they generated. When doing so some interesting trends appear. It can be seen that regardless of posting location, number of replies, and comment label, that the replies all lean toward *Liberal*. In fact, all of the reply chains have more liberal replies then they do conservative. This is despite the conservative corpus having a greater number of words in it. This continues to support the theory that most of the commentary that exists within political arenas online tends to lean toward *Liberal.* This is also despite the number of comments being evenly distributed with both *Conservative* and *Liberal* labels. It appears that once users start participating that the nature of the participation tends to be *Liberal*.

Another finding was the number of times the original commenter appears in the chain. The example discussed previously had the original commenter posting only once to the chain but some examples have upwards of 10 postings by the original commenter. This shows that some users are able to inspire others to reply without having to put in much of an effort, while others are willing to continue to participate in the chain.



Labels of Most Active Influencer (LDA + State of the Union)

Producer	Label2
Fox News	Conservative
Fox News	Nuetral
Fox News	Nuetral
Fox News	Liberal
Fox News	Conservative
Fox News	Conservative
Fox News	Nuetral
Fox News	Nuetral
Fox News	Liberal
Fox News	Liberal

[18]

To see an example of the challenges associated with labeling based on ideology one can observe the commenter that participated often within the chain. We can see from the labeling provided in figure 18 that the user Todd Ambrose, has been assigned both Conservative and Liberal comments. This creates an issue when interpreting because it develops more questions than answers about the nature of the comments being posted. One cannot guarantee the labeling of ideology of the users presented in this research, but seeing an even split on the ideologies does suggest that a more detailed corpus, once developed, could better determine where on the political spectrum this user's comments exist.

As a whole the data set is meant to be an experiment in understanding those who participate online. This is not meant to assure the accuracy of labeling users political ideology, as many users participate few times in the data set, and it is impossible to determine exactly where one stands politically based on a couple comments that were left on a YouTube video. None the less, when one leverages the assumptions stated previously, one can begin to see how with time this approach could develop and become increasingly accurate. The future of such projects would require a detailed corpus that could accurately assign labels by carefully considering all of the words and phrases that can make a post be assigned as *Liberal* or *Conservative*. Such ideas should be discussed in greater detail, as society begins to participate increasingly online there is more that can be understood about what it means for users to exist on various platforms. It has long been assumed that online forums are full of noise and nonsense and this assumption is a reasonable one, as sifting through the data for this analysis made it clear that to find meaning online one needs much more than a hashtag and a couple of key words. One needs a strong understanding of the linguistics needed to understand what exactly people are referring to when they present concepts online.

Discussion

The intent of this research is to better understand the nature of how online commentary can inspire a certain type of discussion. It has so far been found that it often contains harsh critiques and language to match. The nature of comment sections does not promote a respectful discussion between two people who wish to reach a conclusion. It is instead more of a safe space for those who wish to make claims. These claims are not necessarily researched, well developed, productive, considerate, kind, useful, or an accurate representation of the way people believe discourse should occur.

What was found was that the comments that generated the most responses were considerate of grammar, as they contained no spelling errors, used traditional casing, and often contained appropriately used punctuation. These characteristics are not true of all comments and are not true of the replies to many of the examples leveraged for this research. The examples used for this research can be seen below, and they have been orderd by number of responses generated.

1. *Republicans can only create two things: deficits and war."*
2. *"Isn't it funny a socialist is not paying her fair share?"*
3. *"What about the president who wanted our troops home?"*
4. *"John Kerry is having tantrums over the killings of his friends."*
5. *"The dnc, as we know it, will cease to exist by the time Trump leaves office...Thank GOD ALMIGHTY!"*
6. *"The red line was crossed. They shouldn't be surprised; this isn't Obama they're dealing with."*
7. *"Election coming up... Better have a war."*
8. *"Pretty soon the media will have this Iranian general a saint."*
9. *"He embezzled from charity while running for office. Of course we think he has done impeachable offenses, Trump has admitted to it."*
10. *"This action was NOT tyrannical. Salami was a legitimate military target."*

Reading through the most responded to comments allows one to see that there is a spread of what can be interpreted as *Liberal or Conservative*. This makes sense as the original data set leveraged a similar number of comments from each content provider. We also can note that the range in number of replies from the most replied to comment to the least is only 98, suggesting that it tends to take a similar number of users to generate the most replied to comment.

One should take note that because these videos are news clips, that their shelf life is not very long. News clips such as these tend only to provide critical information to the masses for the duration of time that the story is distributed to the public. These videos were sourced from January 2020, before the Corona Virus pandemic took over world news. This is why a strong proportion of the comments are in relation to the death of Qasem Soleimani, as his death was the primary news story at the onset of 2020. Since the Corona Virus has spread across the world, news regarding Soleimani has been almost entirely muted by media outlets, as they instead focus their attention on Corona Virus.

I discuss the shelf-life of such news stories so to better explain the nature of why the comments receive a similar number of replies despite being on different channels, which have different numbers of subscribers. All the videos used for this study were collected during the same time frame, and because they are news videos they cover roughly the same issues, even the nature of the videos are similar. They typically contain an individual, or group of individuals, in suits, talking on to a camera. The differences exist within the language the individuals use when discussing the current events of the day. The commentators who report on behalf on MSNBC are typically considered to be liberal leaning, while those who commentate on Fox are typically associated with a conservative ideology.

What was observed when studying the most replied to comments in the data set was that they typically align with the ideology one would associate with the channel. This makes sense as one would assume that those who are most engaged with the content being provided are likely supporters of that content, or, given this research, that political ideology. The comments also appear to have a similar length, they are all roughly 10 words long and appear as single statements and questions. They do not go into significant detail and do not promote a discussion per say. The statements include accusations of crimes committed and defenses for those who have been accused of wrong doing. I would be inclined to believe that this would further support a belief that those who participate in political arenas online are typically combative and proactive. The intent of these statements appears to promote division amongst those who participate. There is little reason to deduce from these statements that there is a desire to come together, or to give alternative opinions a chance to be heard. Regardless of one's ideology, the nature of these comment sections appears to be one that will continue to create a political divide, at least for those who choose to participate.

It is understood that the current nature of comment forums is combative and often filled with users going back and forth about various issues. The issues discussed are typically inspired by the content of the video. For example, the titles of the videos that contain the two most replied to comments, one from MSNBC and one from Fox News are: *"The Most Seismic Event in the Middle East in The 21st Century"* and *"Alexandria Ocasio-Cortez riles Democrats by refusing to pay party dues".* The first title, which refers to an MSNBC video was a segment from Morning Joe and the second refers to a Fox News video from the program Fox & Friends. Of the two videos the Fox News video generated a larger number of views, by roughly 100,000. Despite the additional views and subscribers, the video generated 29 less replies from its most replied to comment. This appears to suggest that those who watch MSNBC content are more active within the comment section. This may explain why those who are associated with *Liberal* politics often receive the activist label and are often perceived as those who focus a significant amount of their political effort making claims online. It may also stem from liberals being associated with a younger audience that may be more inclined or better skilled at participating online. The increase in views and subscribers may be a result of the age of those who are participating. This assumes that young people are typically liberal and more active online, while older people are typically conservative and less technologically literate. The sustainably greater number of subscribers to Fox News may be a result of a less active, but larger base. A base that chooses to consume content instead of participating in the permanently public forum regarding the content of the video.

Conclusion

This research had the goal of determining if the spaces where individuals post commentary online are inherently segregated by ideology. What was found was that content that has a *Liberal* tendency does contain influential users who actively promote the concepts that would likely be supported by a *Liberal* audience. With regard to *Conservative* content it was found that there is a strong presence of claims that would typically be associated with *Liberal* ideologies. However, observing *Conservative* content revels a much lesser presence of *Liberal* claims than when one observes *Liberal* content.

To confirm that this is that case two corpuses were leveraged to create a mechanism that could label a content as *Liberal* or *Conservative.* Comparing the results of the two labeling strategies revealed that in both cases there is a slightly increase presence of *Liberal* words however the distributions of each ideology remain consistent regardless of approach.

It seems that observing those who are the most influential, on either side of the political spectrum reveals that each of the content provides do provide safe spaces for those of each ideology to make claims that would likely be supported by the content provider. However, it seems that those of the *Liberal* ideology would be associated with a larger space where they can make claims. This is despite *Conservative* content typically having more subscribers and more viewers. Ultimately, while this research focuses only on two of the key players in the United States' political commentary arena it does appear to reveal that the spaces they create are segregated and do little to inspire a discussion or welcome differing opinions or ideas.

The use of the corpus approach leveraged both unsupervised machine learning and previous classification research to create a detailed corpus that could accurately label a comment. However different approaches could push research of this nature forward and reveal, with greater accuracy how prevalent ideological segregating is online. As technology develops and as the supply of data grows research of this nature will become easier to conduct. An approach of this nature is not limited to politics either. Firms could use this approach to understand cliental in an effort to better understand what clusters of their consumers are discussing. Manufacturers could use a similar approach to revel if clusters occur within their employees beyond just where one exists in the employment hierarchy. So long as there are users making claims regarding an understandable number of topics one could use this approach to estimate what type of participation is occurring within forums. Research of this nature need not be limited to big tech firms and it need not be limited to academics. It can be performed with an understanding of algorithm application and data structures. Each of which are gradually being introduced to students earlier in school, as technology continues to develop and become intertwined with our society.

Privacy concerns do arise when looking to understand what users are talking about. However, all the data used for this research was taken from public forums, where the users understand that others will be able to read what they post. It is important that we educate people to become wise users. If someone looks to keep something private, the internet is not the place to look to store it. The technology is designed to be constantly adapting, because the technology is always at risk of being compromised, or used by a graduate student to conduct research. It is important that individuals be especially considerate about what they post online, as the resources being used to study what it being said are improving, and the safety measures being employed often provide some security but are not without flaw. An example to consider would be how messaging apps insist that your messages are encrypted, thus assuring that nobody can access your messages. However, encryption provides no defense for someone screenshotting the messages they shared with you and selling them to the highest bidder.

Ultimately this research has its limitations, most of which stem from the subjective nature of ideology and the wild west nature of online participation. However, what can be understood confirms past research, while also suggesting that not all platforms should be considered equally, as users are assigned different rules and thus can participate differently. The character limitations are much more stringent on Twitter than they are on YouTube or Reddit. None the less it was revealed that the comment sections on YouTube have segregated tendencies and that the tendencies tend to be associated with those who make comments that would typically be considered liberal. I encourage the continuation of this research by leveraging different genres, different size channels, and different countries of reference. Each of these realms could reveal findings that are similar to those found in this research, but they could also reveal differences. The internet, and the content posted to it is remarkably diverse, but those who participate on it do have their similarities. These similarities are what may be able to bridge the divide, and understanding of these differences could inspire people to come together, or drive us further apart. In time one of these realities.

Reference

Andrews, T. M. (2018, December 7). Kevin Hart says he's out as Oscars host after outrage over homophobic tweets. *The Washington Post*. Retrieved from homophobic-tweets-resurface-after-he-is-announced-oscars-host/

Barberá, P. (2015). Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data. *Political Analysis, 23*(1), 76-91. doi:10.1093/pan/mpu011

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*.

Blei, D., Ng, A., & Jordan, M. (2001). *Latent Dirichlet Allocation* (Vol. 3).

Bollen, J., Pepe, A., & Mao, H. (2009). Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *Computing Research Repository - CORR*.

Douiji, Y., Hajar, M., & Hassan, A. (2016). Using YouTube Comments for Text-based Emotion Recognition. *Procedia Computer Science, 83*, 292-299. doi:10.1016/j.procs.2016.04.128

Guyon, I., & Elisseeff, A. (2008). An Introduction to Feature Extraction. In (Vol. 207, pp. 1-25).

Kalra, V., & Agrawal, R. (2018). *Importance of Text Data Preprocessing & Implementation in RapidMiner*.

Politimind. Liberal or conservative Classifier. (n.d.). Retrieved April 5, 2020, from https://uclassify.com/browse/politimind/liberal-or-conservative?input=Text

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., . . . Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Schultes, P., Dorner, V., & Lehner, F. (2013). *Leave a Comment! An In-Depth Analysis of User Comments on YouTube*. Paper presented at the Proceedings of the 11th International Conference on Wirtschaftsinformatik (WI2013), Volume 1, Feb 27 - Mar 01 2013, University Leipzig, Germany. Ed.: R. Alt.

Siersdorfer, S., Chelaru, S., Nejdl, W., & Pedro, J. S. (2010). *How useful are your comments? analyzing and predicting youtube comments and comment ratings*. Paper presented at the Proceedings of the 19th international conference on World wide web, Raleigh, North Carolina, USA. https://doi.org/10.1145/1772690.1772781

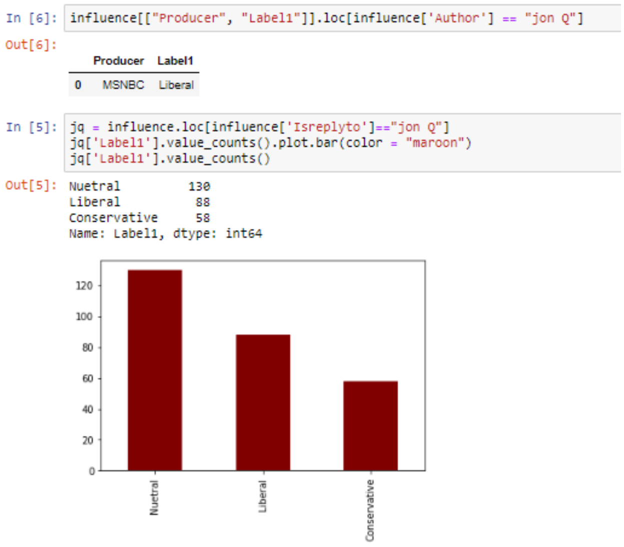
Rob. The Words Democrats and Republicans Use. (2014, February 7). Retrieved April 5, 2020, from https://www.robweir.com/blog/2014/02/the-words-democrats-and-republicans-use.html

Walther, J. B., DeAndrea, D., Kim, J., & Anthony, J. C. (2010). The Influence of Online Comments on Perceptions of Antimarijuana Public Service Announcements on YouTube. *Human Communication Research, 36*(4), 469-492. doi:10.1111/j.1468-2958.2010.01384.x

Webster, M. (2020, February 4). Sentiment. Retrieved February 10, 2020, from https://www.merriam-webster.com/dictionary/sentiment

YouTube. (2020). YouTube for Press. Retrieved February 12, 2020, from https://www.youtube.com/about/press/

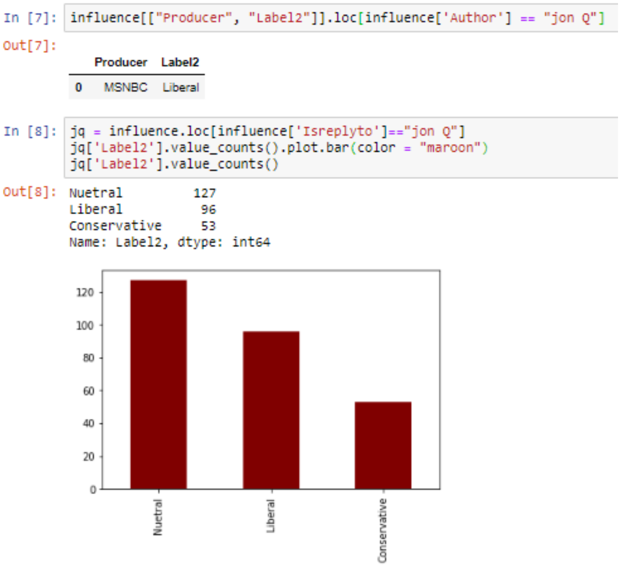
Appendix



```
In [6]: influence[["Producer", "Label1"]].loc[influence['Author'] == "jon Q"]
Out[6]:
```

	Producer	Label1
0	MSNBC	Liberal

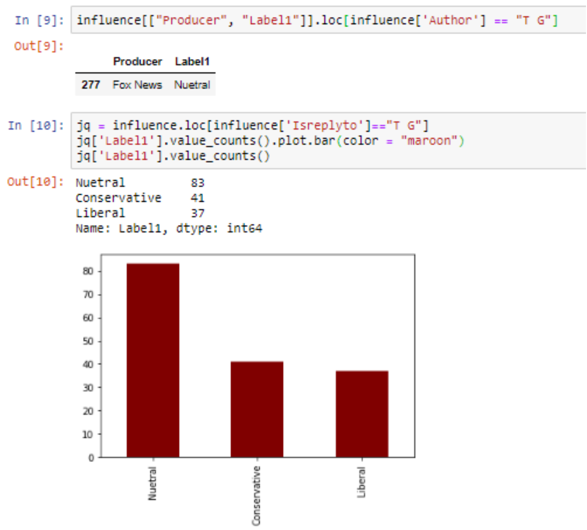
```
In [5]: jq = influence.loc[influence['Isreplyto']=="jon Q"]
        jq['Label1'].value_counts().plot.bar(color = "maroon")
        jq['Label1'].value_counts()

Out[5]: Nuetral         130
        Liberal          88
        Conservative     58
        Name: Label1, dtype: int64
```



```
In [7]: influence[["Producer", "Label2"]].loc[influence['Author'] == "jon Q"]
Out[7]:
```

	Producer	Label2
0	MSNBC	Liberal

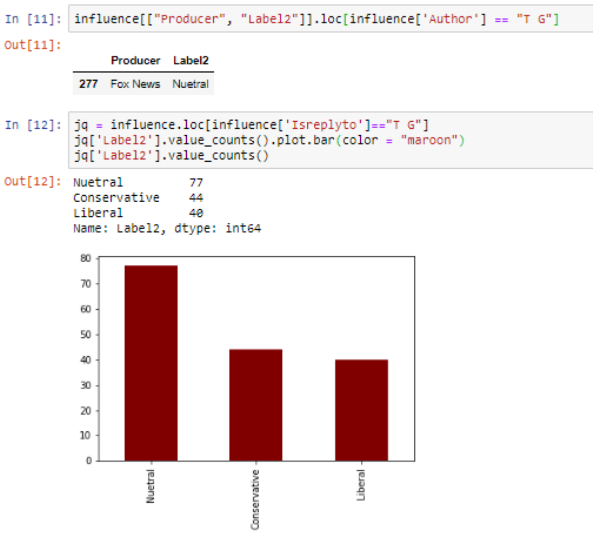
```
In [8]: jq = influence.loc[influence['Isreplyto']=="jon Q"]
        jq['Label2'].value_counts().plot.bar(color = "maroon")
        jq['Label2'].value_counts()

Out[8]: Nuetral         127
        Liberal          96
        Conservative     53
        Name: Label2, dtype: int64
```



```
In [9]: influence[["Producer", "Label1"]].loc[influence['Author'] == "T G"]
Out[9]:
```

	Producer	Label1
277	Fox News	Nuetral

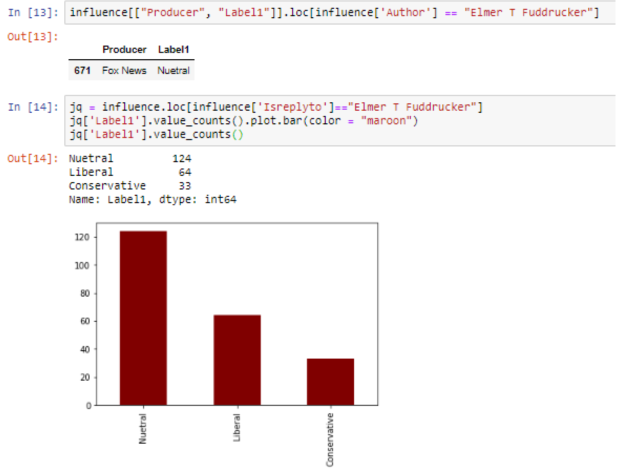
```
In [10]: jq = influence.loc[influence['Isreplyto']=="T G"]
         jq['Label1'].value_counts().plot.bar(color = "maroon")
         jq['Label1'].value_counts()

Out[10]: Nuetral         83
         Conservative    41
         Liberal         37
         Name: Label1, dtype: int64
```



```
In [11]: influence[["Producer", "Label2"]].loc[influence['Author'] == "T G"]
Out[11]:
```

	Producer	Label2
277	Fox News	Nuetral

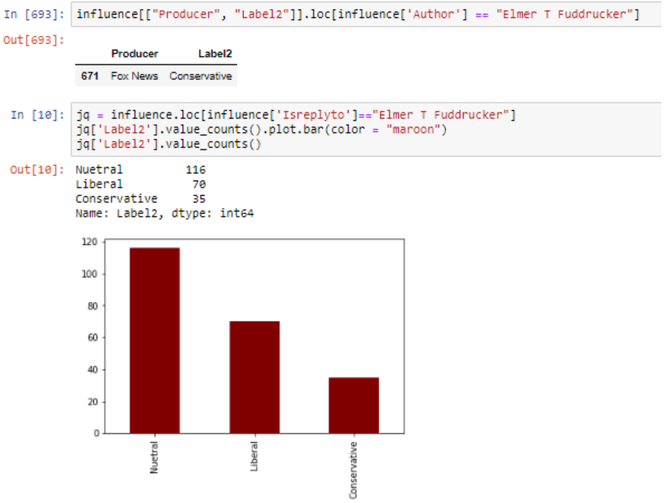
```
In [12]: jq = influence.loc[influence['Isreplyto']=="T G"]
         jq['Label2'].value_counts().plot.bar(color = "maroon")
         jq['Label2'].value_counts()

Out[12]: Nuetral         77
         Conservative    44
         Liberal         40
         Name: Label2, dtype: int64
```



```
In [13]: influence[["Producer", "Label1"]].loc[influence['Author'] == "Elmer T Fuddrucker"]
Out[13]:
```

	Producer	Label1
671	Fox News	Nuetral

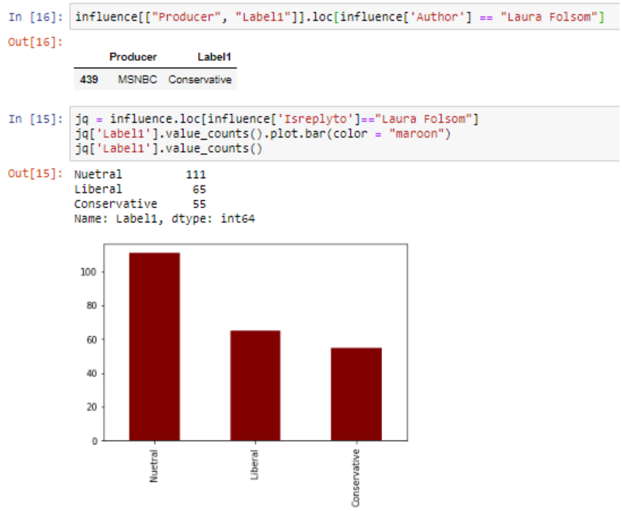
```
In [14]: jq = influence.loc[influence['Isreplyto']=="Elmer T Fuddrucker"]
         jq['Label1'].value_counts().plot.bar(color = "maroon")
         jq['Label1'].value_counts()

Out[14]: Nuetral         124
         Liberal          64
         Conservative     33
         Name: Label1, dtype: int64
```



```
In [693]: influence[["Producer", "Label2"]].loc[influence['Author'] == "Elmer T Fuddrucker"]
Out[693]:
```

	Producer	Label2
671	Fox News	Conservative

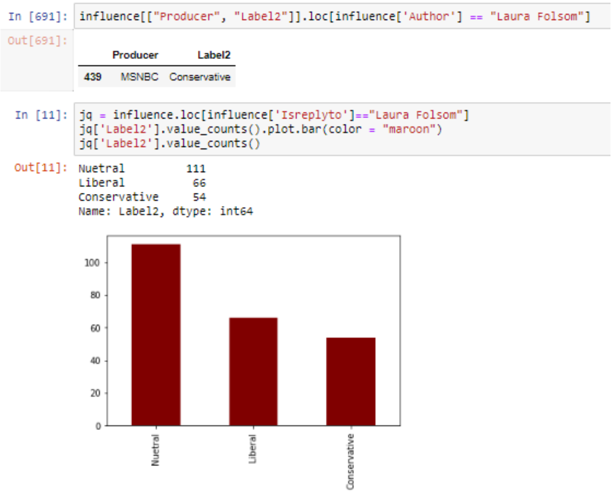
```
In [10]: jq = influence.loc[influence['Isreplyto']=="Elmer T Fuddrucker"]
         jq['Label2'].value_counts().plot.bar(color = "maroon")
         jq['Label2'].value_counts()

Out[10]: Nuetral         116
         Liberal          70
         Conservative     35
         Name: Label2, dtype: int64
```



```
In [16]: influence[["Producer", "Label1"]].loc[influence['Author'] == "Laura Folsom"]
Out[16]:
```

	Producer	Label1
439	MSNBC	Conservative

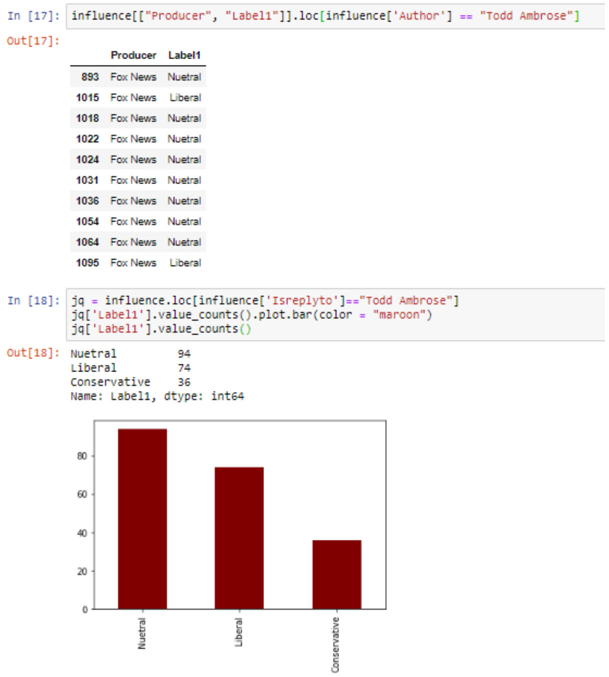
```
In [15]: jq = influence.loc[influence['Isreplyto']=="Laura Folsom"]
         jq['Label1'].value_counts().plot.bar(color = "maroon")
         jq['Label1'].value_counts()

Out[15]: Nuetral         111
         Liberal          65
         Conservative     55
         Name: Label1, dtype: int64
```



```
In [691]: influence[["Producer", "Label2"]].loc[influence['Author'] == "Laura Folsom"]
Out[691]:
```

	Producer	Label2
439	MSNBC	Conservative

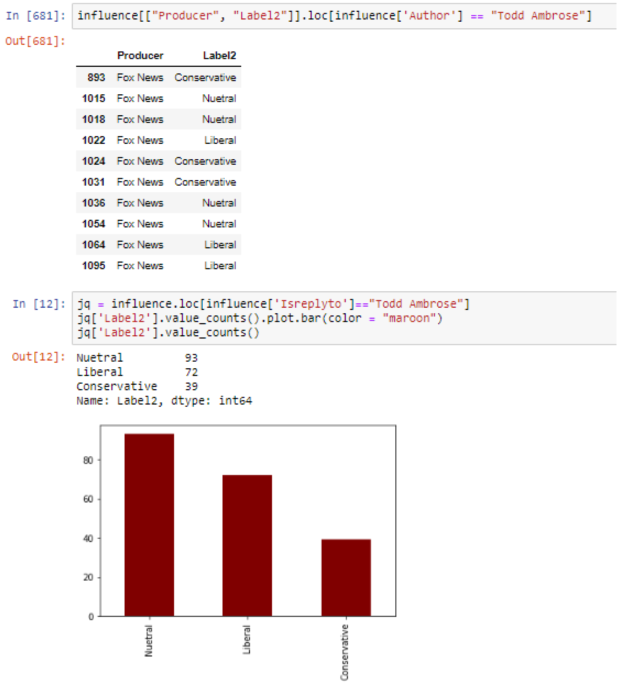
```
In [11]: jq = influence.loc[influence['Isreplyto']=="Laura Folsom"]
         jq['Label2'].value_counts().plot.bar(color = "maroon")
         jq['Label2'].value_counts()

Out[11]: Nuetral         111
         Liberal          66
         Conservative     54
         Name: Label2, dtype: int64
```



```
In [17]: influence[["Producer", "Label1"]].loc[influence['Author'] == "Todd Ambrose"]
Out[17]:
```

	Producer	Label1
893	Fox News	Nuetral
1015	Fox News	Liberal
1018	Fox News	Nuetral
1022	Fox News	Nuetral
1024	Fox News	Nuetral
1031	Fox News	Nuetral
1036	Fox News	Nuetral
1054	Fox News	Nuetral
1064	Fox News	Nuetral
1095	Fox News	Liberal

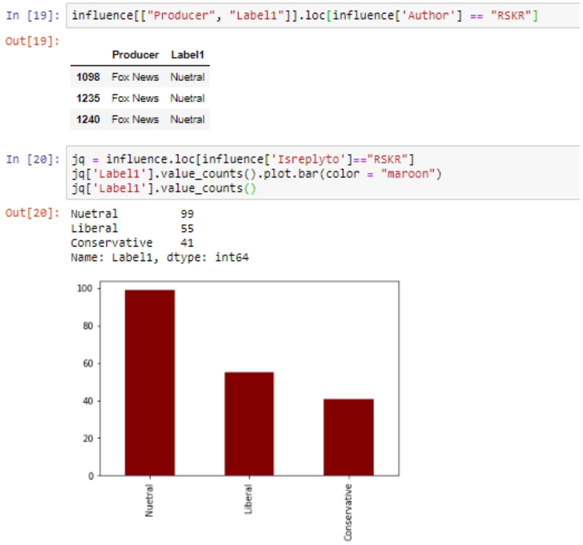
```
In [18]: jq = influence.loc[influence['Isreplyto']=="Todd Ambrose"]
         jq['Label1'].value_counts().plot.bar(color = "maroon")
         jq['Label1'].value_counts()

Out[18]: Nuetral         94
         Liberal         74
         Conservative    36
         Name: Label1, dtype: int64
```



```
In [681]: influence[["Producer", "Label2"]].loc[influence['Author'] == "Todd Ambrose"]
Out[681]:
```

	Producer	Label2
893	Fox News	Conservative
1015	Fox News	Nuetral
1018	Fox News	Nuetral
1022	Fox News	Liberal
1024	Fox News	Conservative
1031	Fox News	Conservative
1036	Fox News	Nuetral
1054	Fox News	Nuetral
1064	Fox News	Liberal
1095	Fox News	Liberal

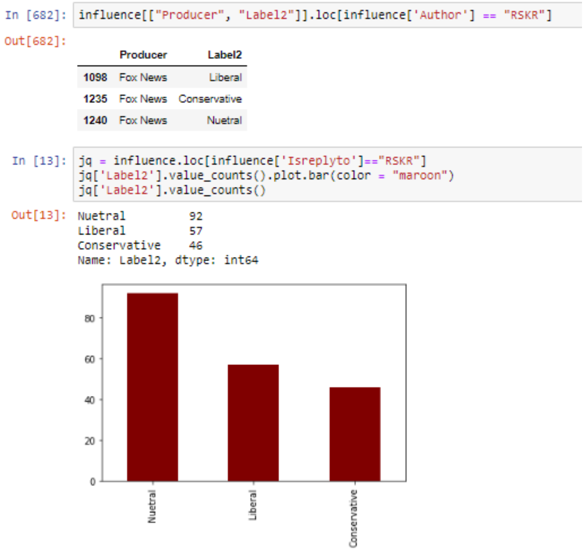
```
In [12]: jq = influence.loc[influence['Isreplyto']=="Todd Ambrose"]
         jq['Label2'].value_counts().plot.bar(color = "maroon")
         jq['Label2'].value_counts()

Out[12]: Nuetral         93
         Liberal         72
         Conservative    39
         Name: Label2, dtype: int64
```



```
In [19]: influence[["Producer", "Label1"]].loc[influence['Author'] == "RSKR"]
Out[19]:
```

	Producer	Label1
1098	Fox News	Nuetral
1235	Fox News	Nuetral
1240	Fox News	Nuetral

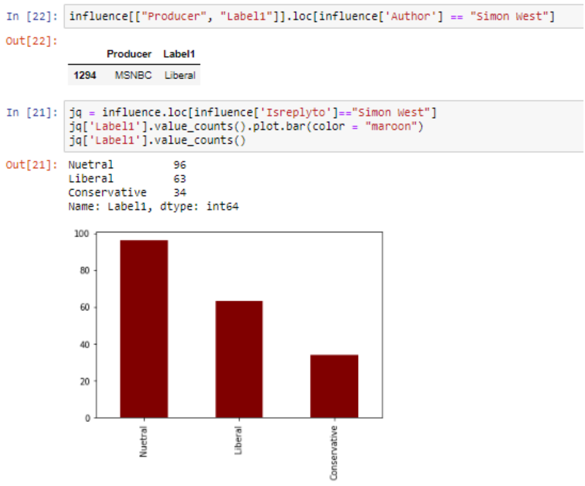
```
In [20]: jq = influence.loc[influence['Isreplyto']=="RSKR"]
         jq['Label1'].value_counts().plot.bar(color = "maroon")
         jq['Label1'].value_counts()

Out[20]: Nuetral         99
         Liberal         55
         Conservative    41
         Name: Label1, dtype: int64
```



```
In [682]: influence[["Producer", "Label2"]].loc[influence['Author'] == "RSKR"]
Out[682]:
```

	Producer	Label2
1098	Fox News	Liberal
1235	Fox News	Conservative
1240	Fox News	Nuetral

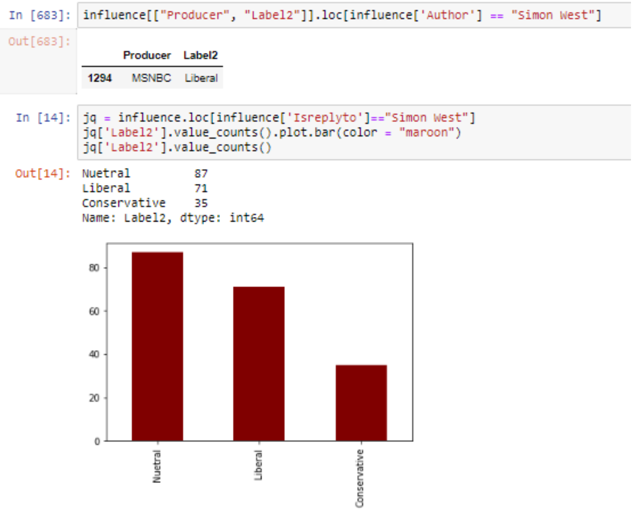
```
In [13]: jq = influence.loc[influence['Isreplyto']=="RSKR"]
         jq['Label2'].value_counts().plot.bar(color = "maroon")
         jq['Label2'].value_counts()

Out[13]: Nuetral         92
         Liberal         57
         Conservative    46
         Name: Label2, dtype: int64
```



```
In [22]: influence[["Producer", "Label1"]].loc[influence['Author'] == "Simon West"]
Out[22]:
```

	Producer	Label1
1294	MSNBC	Liberal

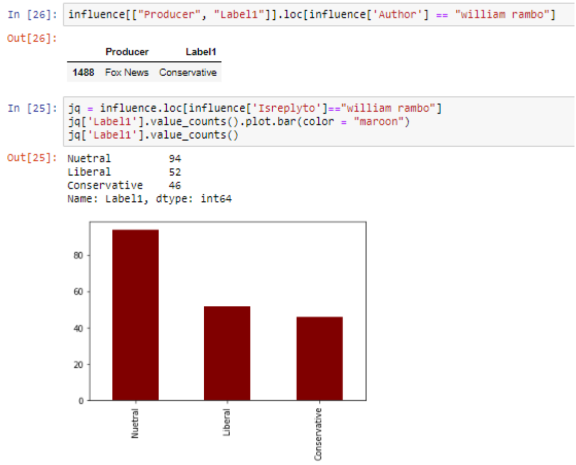
```
In [21]: jq = influence.loc[influence['Isreplyto']=="Simon West"]
         jq['Label1'].value_counts().plot.bar(color = "maroon")
         jq['Label1'].value_counts()
Out[21]: Nuetral         96
         Liberal         63
         Conservative    34
         Name: Label1, dtype: int64
```



```
In [683]: influence[["Producer", "Label2"]].loc[influence['Author'] == "Simon West"]
Out[683]:
```

	Producer	Label2
1294	MSNBC	Liberal

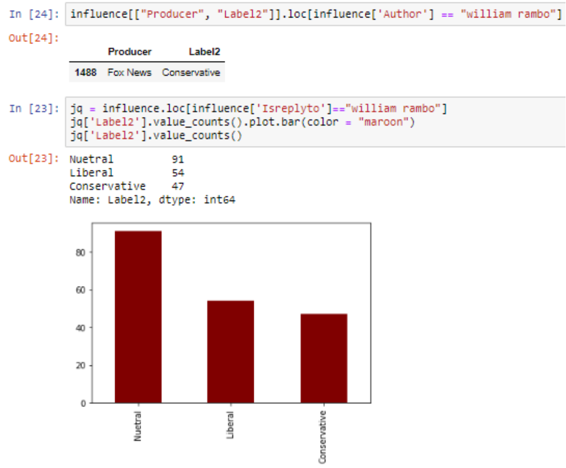
```
In [14]: jq = influence.loc[influence['Isreplyto']=="Simon West"]
         jq['Label2'].value_counts().plot.bar(color = "maroon")
         jq['Label2'].value_counts()
Out[14]: Nuetral         87
         Liberal         71
         Conservative    35
         Name: Label2, dtype: int64
```



```
In [26]: influence[["Producer", "Label1"]].loc[influence['Author'] == "william rambo"]
Out[26]:
```

	Producer	Label1
1488	Fox News	Conservative

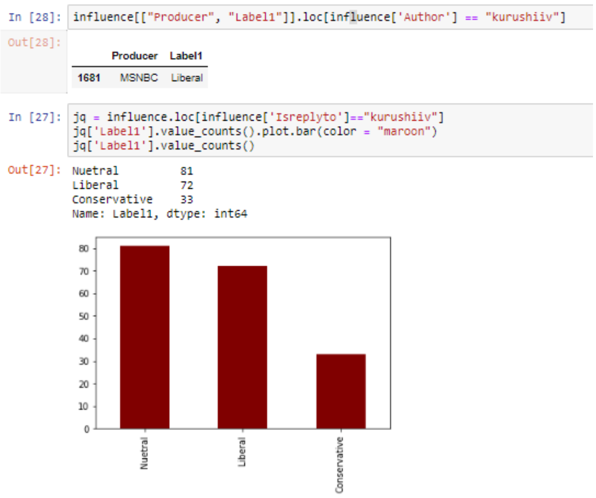
```
In [25]: jq = influence.loc[influence['Isreplyto']=="william rambo"]
         jq['Label1'].value_counts().plot.bar(color = "maroon")
         jq['Label1'].value_counts()
Out[25]: Nuetral         94
         Liberal         52
         Conservative    46
         Name: Label1, dtype: int64
```



```
In [24]: influence[["Producer", "Label2"]].loc[influence['Author'] == "william rambo"]
Out[24]:
```

	Producer	Label2
1488	Fox News	Conservative

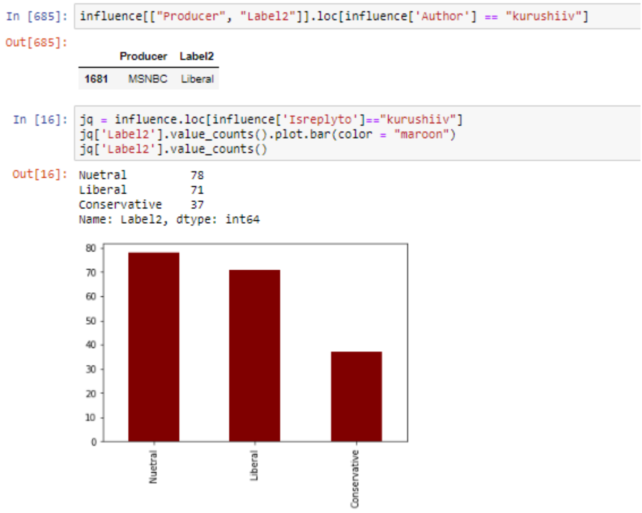
```
In [23]: jq = influence.loc[influence['Isreplyto']=="william rambo"]
         jq['Label2'].value_counts().plot.bar(color = "maroon")
         jq['Label2'].value_counts()
Out[23]: Nuetral         91
         Liberal         54
         Conservative    47
         Name: Label2, dtype: int64
```



```
In [28]: influence[["Producer", "Label1"]].loc[influence['Author'] == "kurushiiv"]
Out[28]:
```

	Producer	Label1
1681	MSNBC	Liberal

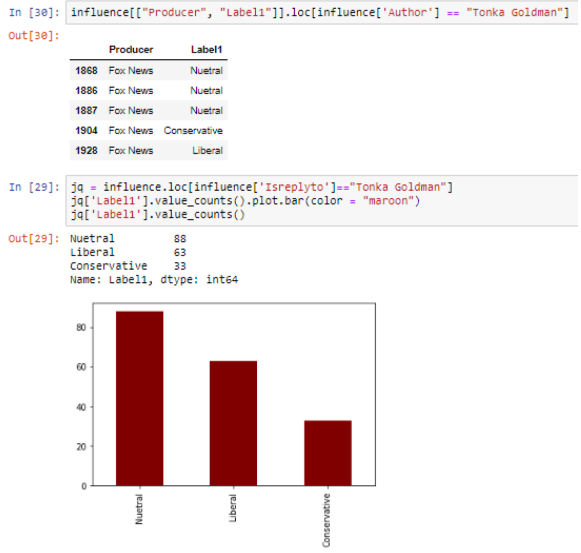
```
In [27]: jq = influence.loc[influence['Isreplyto']=="kurushiiv"]
         jq['Label1'].value_counts().plot.bar(color = "maroon")
         jq['Label1'].value_counts()
Out[27]: Nuetral         81
         Liberal         72
         Conservative    33
         Name: Label1, dtype: int64
```



```
In [685]: influence[["Producer", "Label2"]].loc[influence['Author'] == "kurushiiv"]
Out[685]:
```

	Producer	Label2
1681	MSNBC	Liberal

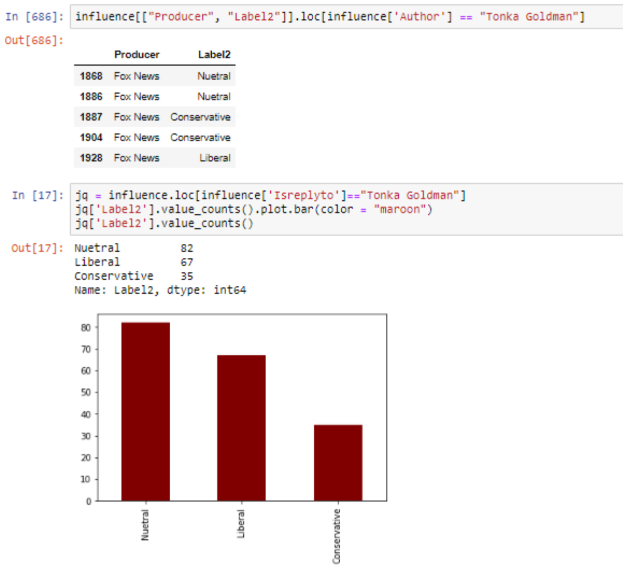
```
In [16]: jq = influence.loc[influence['Isreplyto']=="kurushiiv"]
         jq['Label2'].value_counts().plot.bar(color = "maroon")
         jq['Label2'].value_counts()
Out[16]: Nuetral         78
         Liberal         71
         Conservative    37
         Name: Label2, dtype: int64
```



```
In [30]: influence[["Producer", "Label1"]].loc[influence['Author'] == "Tonka Goldman"]
Out[30]:
```

	Producer	Label1
1868	Fox News	Nuetral
1886	Fox News	Nuetral
1887	Fox News	Nuetral
1904	Fox News	Conservative
1928	Fox News	Liberal

```
In [29]: jq = influence.loc[influence['Isreplyto']=="Tonka Goldman"]
         jq['Label1'].value_counts().plot.bar(color = "maroon")
         jq['Label1'].value_counts()
Out[29]: Nuetral         88
         Liberal         63
         Conservative    33
         Name: Label1, dtype: int64
```



```
In [686]: influence[["Producer", "Label2"]].loc[influence['Author'] == "Tonka Goldman"]
Out[686]:
```

	Producer	Label2
1868	Fox News	Nuetral
1886	Fox News	Nuetral
1887	Fox News	Conservative
1904	Fox News	Conservative
1928	Fox News	Liberal

```
In [17]: jq = influence.loc[influence['Isreplyto']=="Tonka Goldman"]
         jq['Label2'].value_counts().plot.bar(color = "maroon")
         jq['Label2'].value_counts()
Out[17]: Nuetral         82
         Liberal         67
         Conservative    35
         Name: Label2, dtype: int64
```