Bursting the bubble: the role of social bots, millennials, and Trump in an era of political polarization

Sharon Kim, Samantha Stewart, Sanika Bapat

1 Abstract

This paper investigates three different aspects of political activity on Twitter: Trump, political bots, and millennial engagement. To examine Trump as a polarizing agent, we use LDA topic modeling, a classification tree to detect bots, *k*-means clustering, and multivariate linear regressions in our analysis. Tweet entities, specific topic, and personality were statistically significant in predicting user interaction with Trump tweets. Additionally, Trump followers roughly follow Republican demographics. Political bots can also act as a polarizing agent, and can be classified by the ratio of tweets per day to age, followers to friends, and identified by a software called botornot. Based on these factors, it becomes apparent that they follow specific political agendas. To examine how millennials are interacting on Twitter in hopes of dissipating the polarized political arena, the conversations of young, politically active journalists were scraped and visualized using LDA topic modeling. Unfortunately, the conversations show few signs of political convergence and sentiment analyses reveal high levels of strong language among young people, based on sentiment analysis of their conversations.

2 Introduction

Since Donald Trump was elected on November 8th, individuals who otherwise would not engage in a political arena are now encouraged to be involved. Trump uses Twitter to engage directly with his supporters and the public in an unprecedented way. How does Trump use Twitter and how do other users engage with him? Who are his followers, and does that reflect his in person support?

Social media platforms can be easily accessed and used by anyone. This is also true for bots, or automated programs. The research presented here attempts to identify bots that are involved in the current politically turbulent environment caused by the recent 2016 election. The bots will be classified based on 3 factors: the ratio of tweets per day to age, followers to friends and the score provided by botornot. We will also attempt to answer how these bots are contributing to the already-polarized environment and if they have a political agenda. We also want to examine other behavioral characteristics exhibited by these bots in order to improve our classification process for further work in this field.

Moving forward, what tactics can people employ to cultivate a less polarized environment? Given that millennials are the future of politics, we examine their engagement on Twitter. Because public profiles are scant and many other profiles are secure, age is a difficult factor to determine for each user (Nguyen et al., 2013). Instead, we scrape the conversations of young, politically active journalists and

figures from both ends of the political spectrum who are likely to engage with millennials. We collect specifically their conversations because they are more discussion-oriented than other forms of interaction on Twitter, such as retweeting and favoriting (Alvarez-Melis and Saveski, 2016). Additionally, journalists are more likely to cross party lines and follow more candidates (Zhang et al., 2017). Using this data, we visualize the content of the millennials' tweets using LDA topic modeling, and look for topical overlap among the different political ideologies. We also attempt to gauge the open-mindedness of the participants by doing simple sentiment analyses to detect for strong language. With this data, we try to answer the research question: are millennials having genuine discourse that encourages converging political ideologies on Twitter?

3 Literature review

The narrative of political polarization in the United States emerged in the 1990s (Fiorina and Abrams, 2008). Polarization is both a state and a process characterized by the separation of a population into two extremes, leaving few at the center. The framework of political polarization applied to social media is synonymous with "filter bubbles," or like-minded people creating isolated groups. On Twitter, retweet networks containing political hashtags exhibited two major clusters separated by political ideology. The one heterogenous cluster in the mention networks and political hashtags of the opposite political spectrum demonstrated a cross-ideological discussion (Conover et al., 2011). Journalists and users with more followers tended to cross party lines and follow more candidates (Zhang et al., 2017). Additionally, content and context of a tweet can indicate how likely it is to be retweeted. Tweets with URLs, hashtags, and usernames or with emotions had a higher probability of being retweeted. General topics applying to many users, like the economy, were more likely to be retweeted (Naveed et al., 2011).

A socialbot is an automation software that can send users connection requests and post messages on social networks. Previous research has identified bots looking at high ratio of retweets to favorites, and their use of platforms like Botize or Masterfollow (Forelle et al., 2015). Bots avoid discovery by tweeting irrelevant tweets, or "noise," but could be identified if an identical tweet appeared as an independent tweet instead of a retweet (Hegelich and Janetzko, 2016). Zi Chu et al. identify bots using the following factors: (1) levels of entropy, or the number of tweets per unit time; if tweets are periodic (low entropy), the account is likely automated. (2) spam. (3) account properties such as device makeup (Access of Twitter via API indicated bots) and URL ration (ratio of tweets with URLs to tweets without them).

O'Toole et al. argue that the mainstream studies that support the narrative of declining levels of interest in formal politics in young people are flawed because they are limited in their definitions of political participation (O'Toole et al., 2003). Indeed, the opposite trend can be seen as millennials (b. 1985-2004) are more engaged in politics than Generation X was (b. 1965-1985) (Kiesa et al., 2007). One way we hope to see users engaged in politics is through their conversations on Twitter. To this end, Alvarez-Melis and Saveski (2016) propose a more effective model for topic modeling called the conversation pooling method, in which tweets and their replies are aggregated. This addresses the concerns of clustering quality and document retrieval, two factors that posed challenges to previous content-parsing technologies on social media.

4 Data description

4.1 Polarization and Trump

We analyze who engages with Trump on Twitter by looking at his followers. We cannot access all of Trump's 29.4 million followers. Each query returns just fewer than 5,000 accounts. We realized too late that requesting multiple times would return different followers. As such we only have 9,998 unique followers. We assume the low number of followers is not problematic if the API returns accounts randomly. We cannot confirm this because the API is a black box. However, the creation date of the followers accounts ranged from the day of data collection to April 20, 2007, which could indicate randomness.

We use the number of retweets and favorites as indicators of levels of engagement with Trump's tweets. Using the Twitter Rest API and Python package `tweepy` we collected 806 tweets ranging from November 30, 2016 to May 9, 2017. This is acceptable since we are interested in the political implications of Trump's use of Twitter (elected on November 8th, 2016 and inaugurated on January 20th, 2017). We collect tweets twice more to update the counts for number of favorites and retweets. We assume that since Twitter is a platform for instantaneous updates that these counts will stop increasing rapidly after a couple of days, we use 3.

We use Python packages pandas and sklearn to clean and represent the data and statsmodels for our regression analysis. The most difficult part in working with these data were the rate limits set by the Twitter Rest API which reduced the ability to check follow relationships between two users. This meant we could not examine "filter bubbles" or follow relationships between large sets of users.

4.2 Socialbots

We used the Twitter Streaming API to collect tweets which contained the word "Trump", were in reply to Donald Trump, or originated from his account. 100,000 tweets were collected everyday from 4/28 to 5/5. From the 700,000 tweets we were able to isolate 200,000 users. From these users, we were able to identify 250 bots.

4.3 Millennial engagement

We scraped tweet conversations relating to two Democratic political journalists, Lauren Duca (@laurenduca) and Lily Herman (@lkherman), and one Republican figure, Tomi Lahren (@tomilahren). We used the Python package `twitterscraper` to collect the tweets. `twitterscraper` queries Twitter with a variable input and returns a specified number of tweets (n) pertaining to the query. It does not discriminate between tweets; that is, it returns the most recent n number of tweets pertaining to the query. Therefore, we could be sure that we had collected all data related to the query, which was the goal of conversation scraping.

In a search, if a valid user is queried, Twitter returns the activity of the user, which can be broadly defined as the conversations the user is having. More specifically, it can be defined as: all subsequent replies to and including 1) tweets that mention the user, 2) the user's original tweets, and 3) the user's replies to other tweets. It does not return the user's retweets, but it would return the user's reply to the retweet if there existed one. We will be using this definition of a user's activity throughout this paper.

User	Total number of tweets scraped
@lkherman	64,535
@laurenduca	249,365
@tomilahren	312,897

Table 1. We queried 250,000 tweets for each user. Lily Herman's query only returned 64,535 tweets because she is not as active as Lauren Duca or Tomi Lahren on Twitter. We are not quite sure why Tomi Lahren's query returned more tweets than Lauren Duca's query, but the documentation for `twitterscraper` specifies that at least as many tweets as requested would be returned. Although the number of tweets returned is slightly less than the number requested for user Lauren Duca, this difference is irrelevant to our data analysis or results.

5 Methods

5.1 Polarization and Trump

We do two types of analysis to answer the research question. First, we examine who follows Trump using k-means clustering followed by LDA topic modeling. Second, we use multivariate linear regression on Trump's tweet characteristics to predict Twitter user interaction (number of retweets, and number of favorites).

5.1.1 Who follows Trump?

We do unsupervised learning by applying K-means clustering to followers to see if there are specific types of accounts that follow Trump. We suspect that people who voted for Trump, journalists, and republicans would be more likely to follow him on Twitter, and therefore may show up as clusters. We identify binary variables for each. Gender and political orientation are broken up into two dummy variables each. We define an account as fitting one of these groups if their bios contain one or more of the words seen in the table below. Note that it is possible for someone to be genderless and/or have no political orientation.

variable	terms
journalist	journalist, nyt, reporter, media, press
female	mother, mom, daughter, wife, sister, girl, woman
male	father, dad, son, husband, brother, boy, man, guy
democrat	liberal, dem, democrat

republican	republican, libertarian

In this analysis we use age of account since the last data collection, whether the account is verified, the total number of tweets posted from that account (tweet count), follower count, friends count (number of people this account follows), and the ratio of friends to followers. We think this could reflect the social influence of Trump Twitter followers or could pick up bots. However, because the difference in magnitude of variables we normalize across columns. This puts equal weight on each category. We run multiple K-means clustering on this dataset with different values of K. This can be seen in figure **FIGURE 1**.

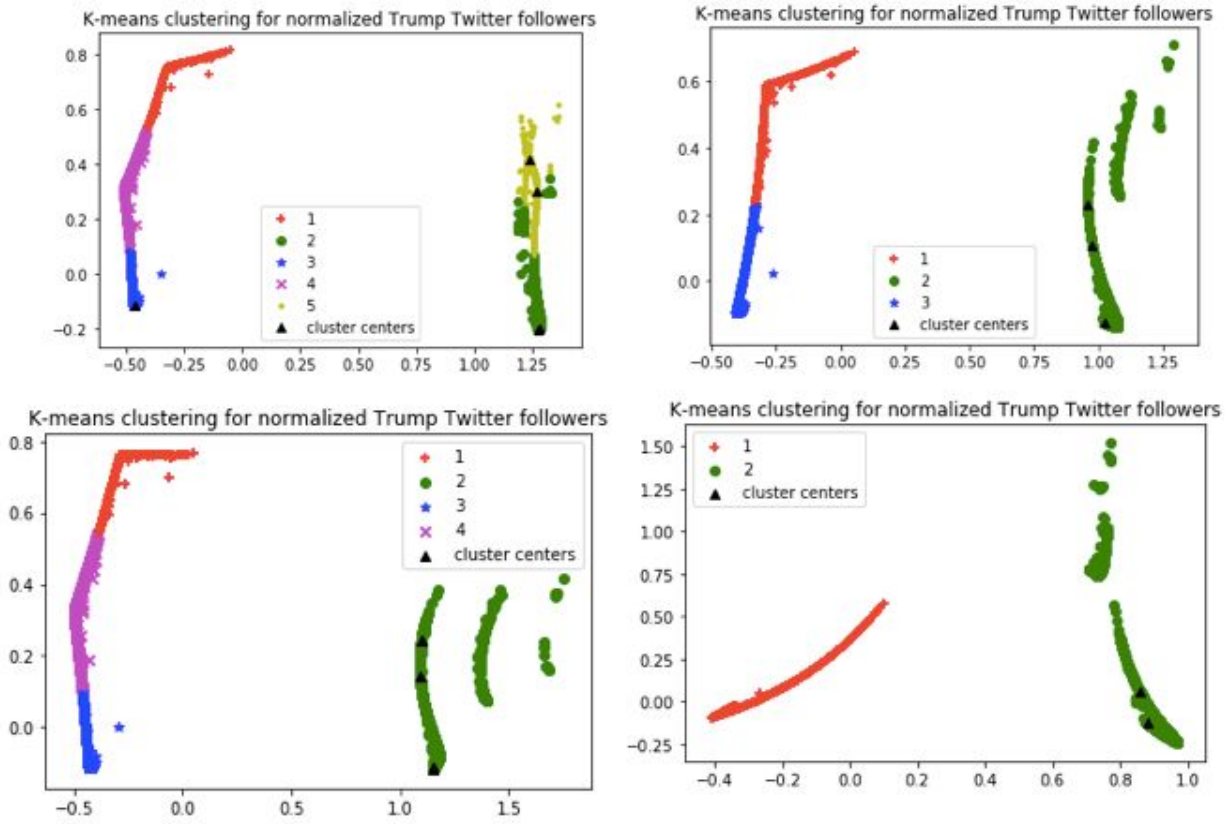


Figure 1 K-means clustering on normalized vectors of Trump followers for K=5 through K=2. Principal component analysis was used to represent this multidimensional data in two-dimensional space.

Graphically, it looks as though two clusters makes a clear divide of those two separate groups. For k-means clustering with 2 clusters, all of the accounts in group 2 (2630), had a description. The top bio was "student" and it occurred 6 times. All of the accounts in group 1 (7358), using the same analysis, did not have descriptions. The difference in follower count, friends count, and age of the account between the two groups can be seen in the chart below.

	Group	50th percentile	75th percentile	100th percentile
Followers count	**Group 1**	1	4	99,902
	Group 2	9	65	493,461
Friends Count	**Group1**	22	63	4,881

	Group 2	91	238	25,602
Age of Account (days)	**Group 1**	52	98.75	3,345
	Group 2	109	1438	d

We refer to how users describe themselves as how they "self-identify." Of the 9,988 followers only 2,630 have non-empty bios. Since K-means clustering split users into two groups based on if they contain a description it does not tell us how they self-identify. We do LDA topic analysis on these existing bios. The output in the following chart and figure are for a tf-idf (term frequency inverse document frequency) LDA model with ten topics.

	Top 15 words
Topic #0:	make, lover, follow, cool, photographer, real, dreams, college, loves, way, company, account, long, gamer, dog
Topic #1:	instagram, snapchat, good, don, know, facebook, com, think, want, people, funny, kind, going, state, stay
Topic #2:	love, god, best, living, media, family, boy, dream, manager, social, loving, friend, children, producer, peace
Topic #3:	like, student, love, things, politics, fun, engineer, guy, government, del, little, happy, india, city, kids
Topic #4:	simple, fan, time, jesus, girl, friends, human, person, head, help, work, team, quality, opinions, dreamer
Topic #5:	just, university, twitter, husband, father, special, hello, travel, king, science, game, lol, christ, video, political
Topic #6:	music, sports, news, football, new, day, wife, business, mother, enjoy, fan, heart, entertainment, trying, great
Topic #7:	world, proud, man, insta, big, years, working, player, director, youtuber, old, senior, marketing, play, tech
Topic #8:	life, live, zorg, love, youtube, voor, management, change, van, doctor, fsu, army, face, project, web
Topic #9:	https, que, por, los, snap, art, para, lifestyle, line, better, sou, musica, amante, official, futbol

Table 4: Across all topics we see spiritual, school-related, relationship-oriented, profession, and social media terms. Topic 9 appears to identify Spanish-speaking users. Topic 1 could indicate younger individuals and topic 4 could be spiritual or religious individuals.

5.1.2 Predicting how users interact with Trump tweets

Other research has looked at "interestingness" on Twitter by using the number of retweets, replies, and favorites as an indicator of interest (Naveed et al., 2011). They found that tweets with hashtags and mentions were more likely to be retweeted. We will use this same idea that the number of retweets and follows can act as a proxy for interest on Twitter. From this we can perform supervised learning by examining which tweet characteristics are correlated with the "success" or increase in interactions with a Trump tweet. This could indicate what users find interesting or important in Trump's tweets.

topic	**terms**
economy	Jobs, economy, employee,
policy	Healthcare, law, bill, congress, senate, tax
election	Hillary, bernie, obamacare, vote, dems, democrats, campaign,

	dnc
Foreign	Russia, china, mexico
conspiracy	Hoax, lies, fake
Personal attacks	Loser, sad, joke

To do this we will run two different sets of multivariate linear regression models with the dependent variable first as retweet count and then again as favorite count. We include variables about the tweet context, whether it was at a late hour (1-5am), and about its content, the number of mentions and hashtags. Additionally we apply counts of topics of interest. The category and terms are seen in the chart right above this. All of the included variables as well as the coefficients, and statistical significance for these regressions are in the following chart.

VARIABLES	(1) retweet	(2) retweet	(3) Retweet	(4) Favorite	(5) Favorite	(6) Favorite
Hashtags(#)	-1935***	-1865**	-1804*	-1.054e+04***	-1.04e+04***	-1.045e+04***
URLs(#)	-5890***	-5426***	-5427***	-2.279e+04***	-1.993e+04***	-2.123e+04***
Mentions(#)	-3067***	-2968***	-3128***	-1.83e+04***	-1.776e+04***	-1.867e+04***
Exclam ("!")		1712**	1725**		1.068e+04***	1.012e+04***
Is Retweet	5450			2.189e+04		
Contains media	-563					
Is late (1-5a)		-556	-686		839	936
Contains "maga"		-1470	-1956	-1305	-4324	-3156
Econ			-454			-1909
policy			-1179			-4076
election			-1765**			-7854***
Foreign			2122*			-2677
Conspiracy			823			3609
Attacks			3246			1.223e+04
Constant	1.597e+04**	2.027e+04**	2.053e+04**	7.283e+04*	8.731e+04**	9.024e+04**
R-squared	0.122	0.128	0.142	0.174	0.189	0.200

*** p<0.01, ** p<0.05, * p<0.1

Figure #: Output for coefficients on OLS multivariate regressions. Dependent variables are column labels.

We find that tweet entities (mentions, hashtags, and URLs) are all statistically significant and negative. This means that for an increase in one of them the number of reweet or favorite counts will decrease by the magnitude of their respective coefficient. We see that Trump's personality, or the number of exclamation marks in his tweets, are statistically significant and positive. Finally, we see that two topics have statistically significant results. For each

additional word that references the election the retweet and follow counts are expected to decrease. For retweets mentioning more policy words actually has a positive effect.

5.2 Socialbots

5.2.1 Identification of Bots

In order to identify bots we examined the ratio of total tweets to the age of the account as well as the followers to friends rations of the accounts. We found that the median for tweets per day was 14 while the 75th percentile was 50. For the ratio of followers to friends the mean was at 1.5, i.e: the user followed 1.5 times the number of people who followed him. The 75th percentile for this ratio was 4.3. Figure 1 further shows how extreme the data set was and illustrates that the bots chosen were extreme outliers.

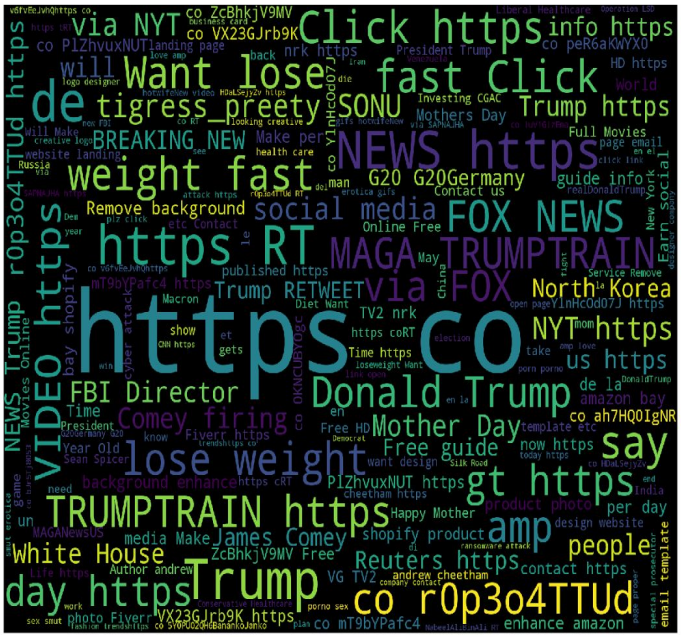
	50th percentile	75th percentile	100th percentile
Tweets per day	14	50	3000
Followers to Friends	1.5	4.3	678

Any accounts that were over the 75th percentile for both of the above categories were labeled as ‘probable bots’. Through this method, we were able to identify 3000 probable bots. These probable bots were then passed through an API for ‘Bot or Not’, which returns a score for the likelihood that an account is a bot. Any account with a score greater than 0.6 was labeled as ‘Bot’. After removal of double accounts, we were able to identify 250 bot accounts.

We then downloaded the latest 200 tweets of each of these bots and pre-processed all tweets by removing any characters that could not be decoded into ASCII (non latin words and emoticons). We also removed any stop words or punctuations and transferred all the tweets to lower case. The resulting document matrix was then subjected to the following statistical analyses:

5.2.2 Descriptive Text Mining

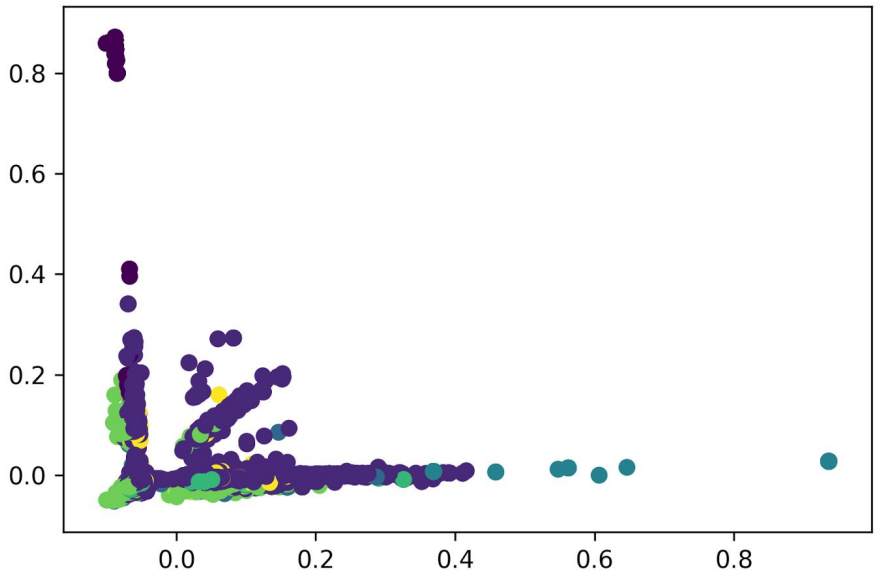
Figure 2 shows a word cloud of the corpus. It’s interesting to see that the most frequent words are either of political importance or of weight loss advertisements. The fact that ‘https’ is the most common word shows that the bots are likely promoting various other websites and URLS. G20 and Germany also show up quite commonly as well as ‘white house’, “Fox News”, and “North Korea”.



A word frequency analyzation showed that only the word ‘Trump’ appeared more that 2000 time in the corpus, while no other words appeared more that a 1000 times.

5.2.3 *k*-means clustering analysis

Next is a description of the clusters and their corresponding tweets revealed by the K-Means Clustering Algorithm. After repeated experimentation we saw that 15 - 20 clusters was the number that seemed to give the most distinct and useful topics.



Cluster 1(601), 3 (349), 4(155) , 5(51), 10(715), ll(1193) and 12(751) consisted of tweets that could not be associated to a specific topic. These accounted for 6.9% of the total tweets in the corpus and were all part of the noise that we expected to see.

Cluster 2(36182) tweets which consisted of almost 77% of the total tweets in the corpus consisted. Typical tweets in this corpus were:

"rt @ed_mealsupport: a powerful message from a lady in eating disorder recovery. #eatingdisorders #anorexia #mentalillness #recovery"

The words appearing most often in this cluster are 'free', 'may', 'love' and 'may', however none of these words are more than 2% of this cluster. Thus we inferred that the primary factor that these tweets have in common is the constantly changing content. This cluster is the primary reason why bots are so hard to identify. Their purpose can be assumed to be two-fold: camouflaging as normal users and presenting content that may be of interest to normal users in order to gain more traction.

Cluster 6(199) consisted of 0.4% of the corpus. Most of the tweets in this cluster were about design and creativity. Typical tweets were :

Cluster 7(1559), which constituted 3% of the corpus, was compiled of Spanish tweets but without any particular agenda. They seemed to be generating noise as the most common word consisted of 'video', 'popular' and 'Madrid'.

Cluster 8(4042) which was 8.6% of the corpus was the most important cluster. This cluster had primarily tweets that were about Donald Trump. 'Trump' was part of 50% of this cluster, along with 'Comey', 'firing', 'Fox', 'FBI' and 'Russia'. This was also the second largest cluster, therefore it can be concluded that though bots generate a significant amount of noise, they are also generating a significant amount of political content. Since Fox news appears often in the cluster, we can also infer that the bots are tweeting right-wing propaganda.

Cluster 9(154) consisted of tweets primarily concerned with European News as 'euro_news' was retweeted most often. Most common were like :

#news| theresa may is much more popular than the conservatives, poll suggests

The words 'poll' and 'support' were also very common.

Cluster 13(411) consisted of tweets primarily concerned with wishes and death. The particular tweet was repeated multiple times:

rt @tigress_preety: sonu
you're every shooting star
every 11:11 wish

Cluster 14(200) consisted of German tweets regarding the G20. Typical tweets were :

#g20 #g20germany - g20 in hamburg - alternatives medienzentrum startet akkreditierung - https://t.co/no4gxpokpy https://t.co/1lkxmfiw3k

The words 'wish' , 'dying' and 'ignore' were very common.

Cluster 15(400) consisted of a single tweet appearing 400 times:

#fat want to lose weight fast? click here https://t.co/vx23gjrb9k https://t.co/fpdkzzckwo

This was probably a message promoted by bots in order to appeal to a larger audience and increase their influence.

LDA analysis

The LDA analysis further affirms the conclusions from the *k*-means clustering. 6 out of the 9 topics contained tweets about Trump and Comey, which shows that the bots have a political agenda. The important topics identified by the LDA analysis are as follows:

Topic	Keywords
1	https, rt, trump, nhttps, new, amp, la, news, 2017, 10, people, comey, maga, want, make
3	https, rt, trump, new, nhttps, amp, la, news, 2017, comey, video, 10, want, people, love
5	https, rt, trump, nhttps, new, amp, comey, news, la, 2017, video, want, just, 10, maga
6	https, rt, trump, nhttps, maga, comey, 2017, news, make, amp, new, la, people, world, 10
8	https, rt, trump, new, nhttps, amp, comey, 10, 2017, en, la, life, video, make, just
9	https, rt, trump, nhttps, new, comey, amp, 10, 2017, la, want, says, video, world, news

5.3 Millennial engagement

Each of the scraped tweets was returned as a tweet object and stored in a text file. Because we aimed to look at the conversations of each user so we could perform LDA topic analysis on the data later, we had to write an additional script to parse through all of their tweets and determine to which conversation each of the tweets belonged. We accomplished this task using the `urllib2` and `BeautifulSoup` libraries, and stored each conversation as a separate JSON file.

For each of the users, it was important to separate tweets concerning her personal life from the politically charged ones. We accomplished this by restricting analysis of her tweets to certain time frames that generated the most amount of activity, assuming those discussions would be most politically dominated. Using the Python package `pandas`, we determined that this time interval hovered around January-April 2017 for each user.

After performing LDA analyses for each of the users, we did a brief sentiment analysis for users Lauren Duca and Tomi Lahren to detect for strong language. We did not do this for user Lily Herman because her LDA results showed very little strong language. We assumed that a higher incidence of strong language would mean an audience that was less interested in having cross-ideological discussions that would push a convergence in ideology.

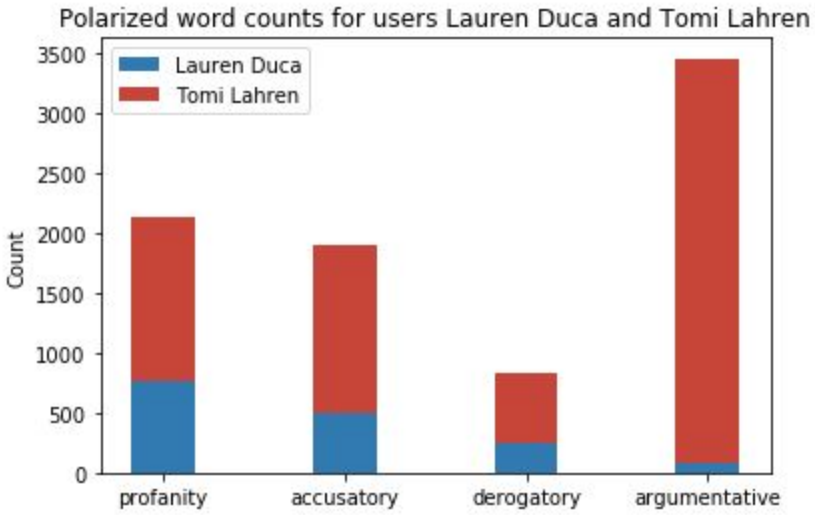


Figure x. Because the ratio of the total number of tweets of Tomi Lahren to Lauren Duca ≅ 1.918, Tomi Lahren's count values were normalized. Examples of accusatory, derogatory, and argumentative words: ["wrong", "hypocrite"], ["snowflake", "stupid"], ["sue", "fire"]. We assume examples of profanity are obvious.

5.3.1 Lily Herman (@lkherman)

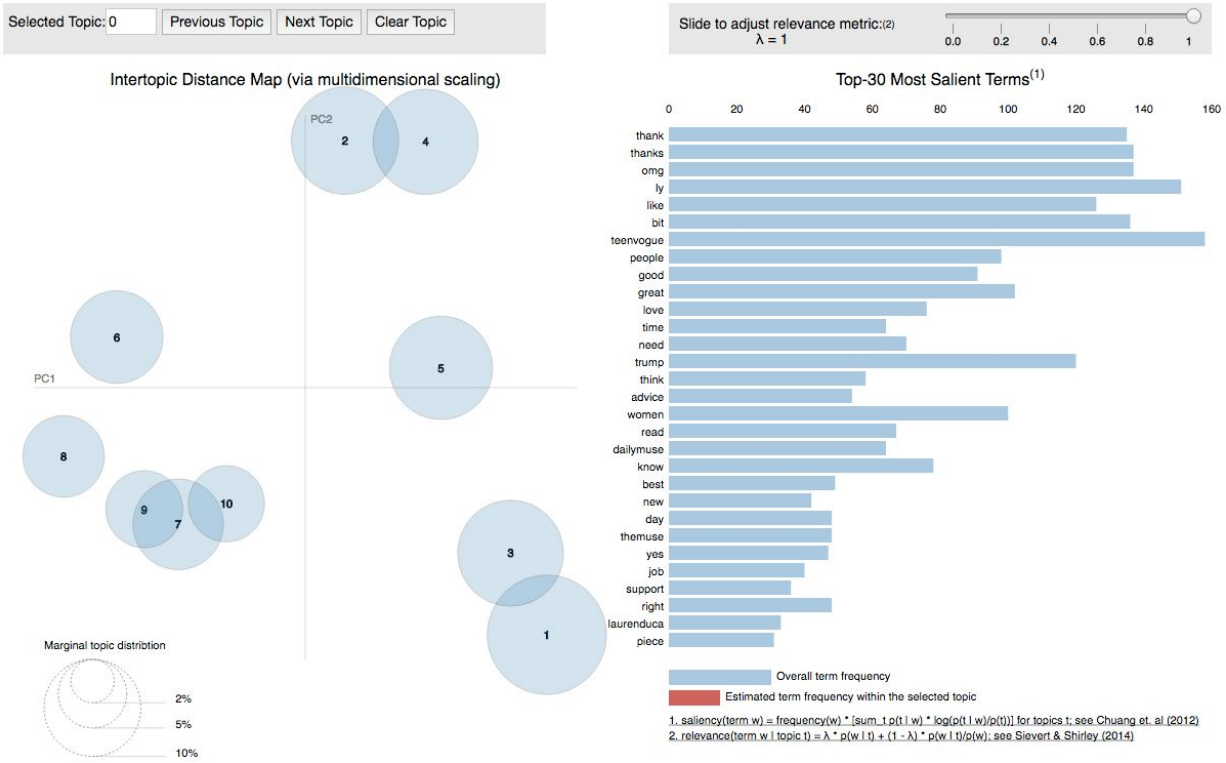


Figure 2. LDA topic model visualization of Lily Herman's activity, January-April 2017. [interactive]

Rather than tweeting about specific political events, Herman's political activity gravitated towards addressing broad issues such as feminism and women in the workplace.

5.3.2 Lauren Duca (@laurenduca)

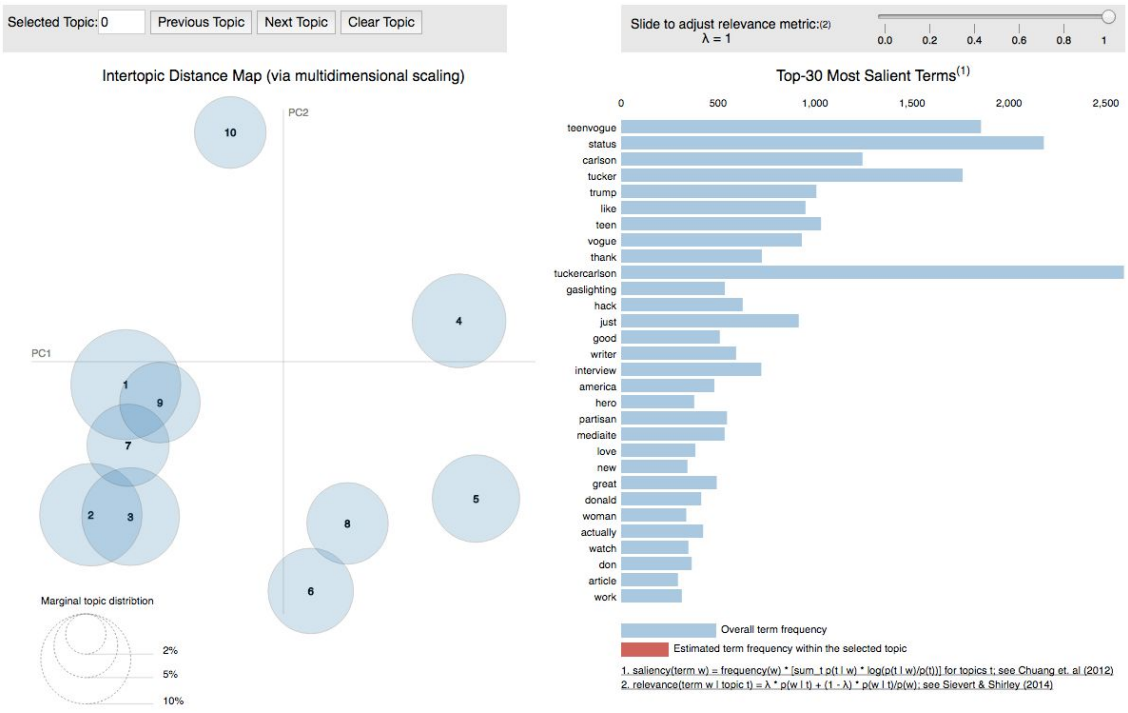


Figure x. LDA topic model visualization of Lauren Duca's activity, November-March 2017. [interactive]

Many of Lauren Duca's topical circles dealt with events such as her recent criticisms of Ivanka Trump and Republican political commentator Tucker Carlson's response to these criticisms and her recent article in TeenVogue magazine about Trump and gaslighting (a term she uses for falsifying information).

5.3.3 Tomi Lahren (@tomilahren)

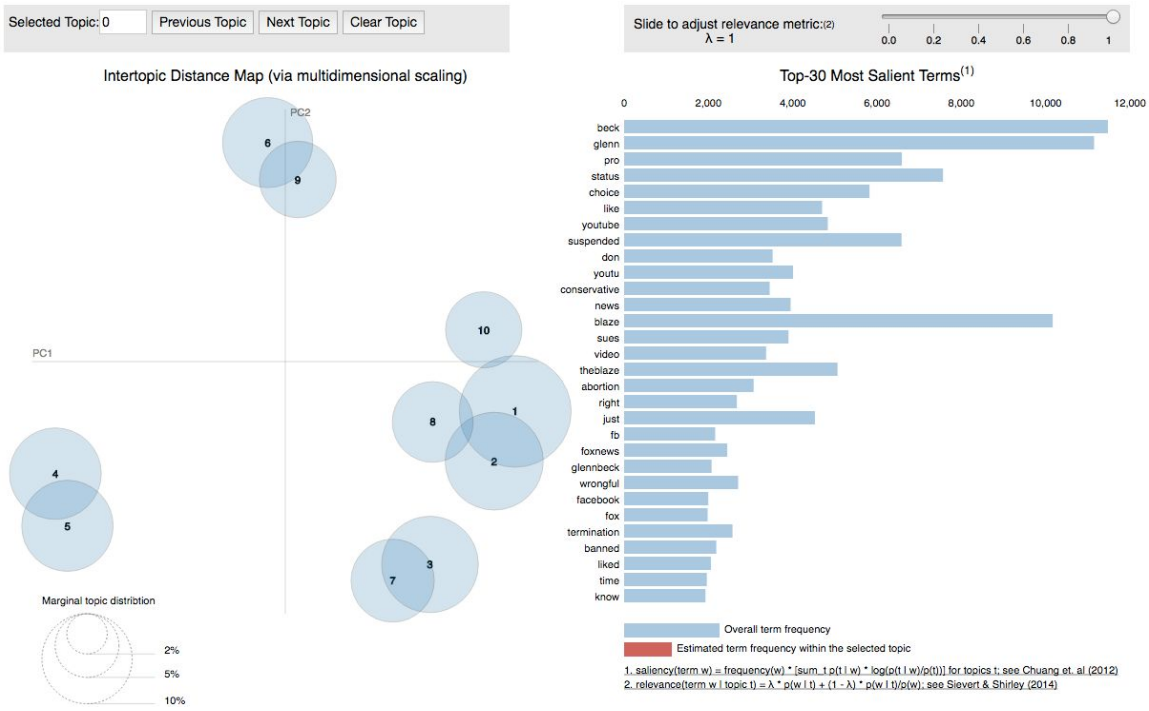


Figure x. LDA topic model visualization of Tomi Lahren's activity, January-April 2017. [interactive]

Tomi Lahren's activity also deals with specific political events, such as her recent firing from The Blaze, a conservative television network, by Republican television host Glenn Beck, for voicing her pro-choice views concerning abortion.

6 Discussion

6.1 Polarization and Trump

6.1.1 Who follows Trump?

We see many of Trump's followers do not have bios and were created recently. It would be interesting to compare this to Twitter at large. We found more male than female users which could support our hypothesis that voting demographics carry over to Twitter. However it is possible that there are more males on Twitter than female or that they self-identify more often. We found fewer journalists than expected. Many followers had low social influence (low follower counts). LDA modeling roughly reflected Republican ideals of family and religion. It also indicated a lot of international followers, specifically Spanish-speakers. This was difficult on short documents especially because the terms in bios tend to be lists of identity traits, not necessarily things that are related. Trump has a large number of

followers and it would not be logical to think that the ones we sampled would be the most interesting of them. Running this analysis on a larger dataset could lead to better results.

6.1.2 Predicting how users interact with Trump's tweets

We found highly statistically significant results for entities in tweets for both retweet and favorite counts. This relationship is negative however, unlike that found in Naveed et al.'s research. Additionally, we see fewer interactions for tweets mentioning the election. It may be the case that Twitter users, whether they follow Trump or not, find Trump's original content more interesting than generic tags. This would make sense because tweets are limited to 140 characters. This would imply that people think his content is interesting. This may also reflect the purpose of retweeting is to spread information. This can either be by individuals supporting Trump's statements or ridiculing them.

This answers our question that tweets are more well-received (favorite count) and interacted with more (retweet count) if it is Trump's original content. This strengthens his argument of using Twitter to connect directly with people.

6.2 Socialbots

The initial question that we wanted to answer where the social bots were concerned were the following:

a) Do the twitter bots have a political agenda?
b) What kind of behaviours do they exhibit?

The analysis shown in the research presented answers to both of the research questions. The twitter bots have a political agenda. Even though the vast majority (77%) of tweets generated by the political bots were nonsense, political themes did emerge (as seen in the LDA analysis). The bots also seemed to exhibit the following traits:

a) Camouflaging their identity by tweeting noise
b) The noise tweeted would be of interest to normal user (tips on weight loss, or free goods)

Both clustering approaches used also lead to comparable results. The K-means clustering was useful in determining the one big cluster of noise, but also the second largest cluster of politics. The LDA analysis was useful in illustration the dominance of political themes (FBI, Trump and Comey) throughout a majority of the topics generated.

The bots that we were able to identify were most definitely not human. However, due to the strict criteria used to identify these bots, it is very likely that we missed quite a few of the smarter bots. However this was a sacrifice that needed to be done in order to not compromise on the accuracy of our results. A supervised learning approach to the process of identification of bots might lead to finding more hidden bots and also uncovering a bot-network.

6.3 Millennial Engagement

The LDA visualizations revealed specific political topics that the users were discussing, but did not show a convergence in political ideology or common themes between each of the users in their discussions.

Based on the simple sentiment analyses of the more politically-engaged Lauren Duca and Tomi Lahren, Tomi Lahren uses higher levels of inflammatory language than does Lauren Duca. In general, however, millennials do not seem to have a problem using strong language to bolster their political ideology, which is counterproductive to open-minded and genuine political dialogue.

Upon reflection, replying to people in under 140 characters does not allow for genuine political discourse. Twitter is filled with sound bites that are intended to generate a lot of attention in the form of likes and retweets. This makes it hard to tell whether the millennials are genuinely interested in having open-minded political discourse, which is probably unlikely.

7 Conclusion

This paper uses the motivation of the current political climate in the United States to investigate trends of political polarization and millennial engagement through scraping Twitter data. We also look at how users on this site interact, particularly with Trump. We investigate socialbot activity because of the increased role of these accounts in political cycles, and scrape millennials' accounts to see if the younger demographic is contributing to dissipating the polarized political atmosphere.

We used a combination of the Twitter Streaming API, Rest API, and web scraping with BeautifulSoup to collect account information, Trump tweet information, Trump followers, and conversations with millennial political journalists. We used exploratory data analysis to understand trends and characteristics of the data. We applied supervised learning in the form of regressions, and unsupervised learning through LDA and *k*-means clustering.

We identified characteristics of Trump's tweets which significantly impacted how they were interacted with and did not follow other research trends. Generally, Trump's online followers, who self-identified in bios, reflect his base of support offline too. Political bots can also act as a polarizing agent, and can be classified by the ratio of tweets per day to age, followers to friends, and identified by a software called botornot. Unfortunately, millennials using Twitter show little signs of engaging in open-minded dialogue that encourages converging political ideologies.

For future studies, we could look at the spread of Trump-originating information through follow networks and see the difference in spread through supporters and opponents. Future research could look into how to block bots' influence. It would be helpful to scrape dialogue from social platforms that allow more lengthy text if we wish to examine more genuine and open-minded political discourse.

8 References

1. Fiorina, M., & Abrams, S. (2008). Political Polarization in the American Public. Annual Review of Political Science, 11, 563-588.

2. Conover, M., Ratkiewicz, J., Francisco, M. R., Gonçalves, B., Menczer, F., & Flammini, A. (2011). Political Polarization on Twitter. ICWSM, 133, 89-96.

3. Naveed, N., Gottron, T., Kunegis, J., & Alhadi, A. C. (2011, June). Bad news travel fast: A content-based analysis of interestingness on twitter. In Proceedings of the 3rd International Web Science Conference (p. 8). ACM.

4. Forelle, M. C., Howard, P. N., Monroy-Hernández, A., & Savage, S. (2015). Political bots and the manipulation of public opinion in Venezuela.

5. Hegelich, S., & Janetzko, D. (2016, March). Are social bots on twitter political actors? Empirical evidence from a Ukrainian social botnet. In Tenth International AAAI Conference on Web and Social Media.

6. Chu, Z., Gianvecchio, S., Wang, H., & Jajodia, S. (2012). Detecting automation of twitter accounts: Are you a human, bot, or cyborg?. IEEE Transactions on Dependable and Secure Computing, 9(6), 811-824.

7. Kiesa, A., Orlowski, A. P., Levine, P., Both, D., Kirby, E. H., Lopez, M. H., & Marcelo, K. B. (2007). Millennials Talk Politics: A Study of College Student Political Engagement. *Center for Information and Research on Civic Learning and Engagement (CIRCLE)*.

8. O'Toole, T., Lister, M., Marsh, D., Jones, S., & McDonagh, A. (2003). Tuning out or left out? Participation and non-participation among young people. Contemporary politics, 9(1), 45-61.

9. Nguyen, D. P., Gravel, R., Trieschnigg, R. B., & Meder, T. (2013). " How old do you think I am?" A study of language and age in Twitter.

10. Alvarez-Melis, D., & Saveski, M. (2016, March). Topic Modeling in Twitter: Aggregating Tweets by Conversations. In ICWSM (pp. 519-522).