All of Trump's Words: Linguistic Change in Online Political Discourse

Alberto Melgoza and Chet Gutwein, UC Berkeley

Abstract

President Trump has used buzzwords and provocative text to "weaponize" his language. The press and social media response, whether in support or opposition, has used his language and repeatedly echoed his words in print which has given him an incredibly powerful marketing tool. We have utilized Snapshot Language Models (SLM) to analyze the impact of Trump's use of language in online communities. Our findings demonstrate that President Trump has impacted our communication patterns in a big way.

1 Introduction

American democracy is currently undergoing the biggest test most of have seen in our lifetimes. This is not in small measure due to the current Presidents sustained disregard for truth and disdain towards people and institutions that enable our democracy to function, such as the free press.

But how does he plan to get away with it? Professor emeritus of linguistics at UC Berkeley Lakoff (2016) explains, that Trump, as the master salesman, has turned words into weapons to manipulate media and public opinion in order to sell whatever narrative serves his goals, regardless of what is true or best for the American people. Professor Lakoff argues that "By faithfully transmitting Trumps words and ideas, the press helps him to attack, and thereby control, the press itself". We believe that, although this argument makes intuitive sense, having quantifiable evidence supporting it would make it that much stronger.

From previous work done with Snapshot Language Models (SLMs) (Danescu-Niculescu-Mizil et al, 2013), we understand that there is a somewhat predictable pattern that users in an online discussion forum typically follow that can be tracked by linguistic change, which Danescu-Niculescu-Mizil et al (2013), defines as: "linguistic innovation originating in a sub-group that becomes accepted as the norm through a process of conforming." We study how much of this linguistic change can be correlated to and impacted by the President's influence.

Automated content on the internet and various social media outlets has become seemingly growing in presence in online discourse. Applications for bots range from advertising products to influencing voters in elections. The latter is considered by the most as the biggest threat that bots pose, and there is evidence that our language patterns through online mediums are increasingly affected by polluted content generated by bots. Our intent was to focus some analysis on bot interaction with language patterns, but instead we are able to point out some evidence that may motivate continued study.

Our Approach. We have chosen the Reddit r/politics subreddit (discussion forum) as the platform for our analysis. We developed SLMs for the time period covering January 2015 through December 2017 (2016 Election) and January 2011 through December 2013 (2012 Election). We tested each SLM's proximity to Trump's language using a set of posts from 2018 that explicitly contain Trumps weapon words such as "fake news", "witch hunt", "deep state". We made direct comparisons between the 2016 Election SLM results and the 2012 Election SLM results.

We made a few departures from the Danescu-Niculescu-Mizil et al (2013) SLM model structure in order to suit the focus of this study. Primary differences with our SLMs are that each SLM is a tri-gram model and we used Kneser-Ney smoothing (Chen and Goodman, 1998) instead of Katz. We utilized perplexity as the pri-

mary measure to evaluate individual posted content and global community-level trends in language used over time.

Summary of Main Findings. Using the metric of perplexity, we have strong evidence to suggest that Trump's use of weapon words has engaged political discussion in a way that has affected our communication patterns in a big way. For any pair of years (2011 vs. 2015, 2013 vs. 2015, etc.) between the two different elections, there is a measureable difference in the perplexity of the test set of 2018 posts. In Figure 1, we can see the immediate impact of Trump weapon words on post activity. The increase of post frequency including usage of weapon words supports Lakoff's theory – we (i.e. the community) are taking his charged and targeted linguistic terms and adopting them into our own speech patterns. Analysis of our test set using SLMs further demonstrates that political discourse has made a shift towards the language that Trump uses. Figure 2 shows perplexity, by post, as an average for SLMs over a full year. Comparisons demonstrate a noticeable, and consistent, difference between perplexity scores for each post.

2 Background

2.1 Snapshot Language Models

Tracking linguistic change in online communities has been performed using bigram Snapshot Language Models (SLM) as demonstrated by (Danescu-Niculescu-Mizil et al, 2013). We have used the same technique and expanded its use with an emphasis on community level linguistic change within online political forums. Within the timeframe specified we have developed a trigram language model using Kneser-Ney smoothing for each month. Previous work has revealed that linguistic change occurs frequently and user behavior patterns are so strongly linked that user lifecycles can be predicted. We wondered what happens when enough users are impacted by the same *influential event*. In such situations, we posit that entire communities use of language can be shifted towards language surrounding an event. The SLMs we generated follow the same framework as (Danescu-Niculescu-Mizil et al, 2013) where a separate SLM_m is generated for every month (ie. SLM_{apr15}, SLM_{may15}, etc.).

Our application of SLMs is geared towards community-level changes and not user-specific behavior, so in this sense our use of SLMs is quite different. Our evaluation of each SLM uses a test set of 2018 posts. Each post, p, is evaluated by calculating *Perplexity* where:

$$2^{H(p,SLM_{m(p)})} = 2^{-\frac{1}{N}\sum logP_{SLM_{m(p)}}(t_i)}$$

where t_1,...,t_N are tri-grams making up text of each post, p. A lower value for perplexity indicates a post that is closer in agreement to $SLM_{m(p)}$. While *perplexity* is typically used as a measure of accuracy of a language model, in our study we are measuring the accuracy in terms of a specific group of words (i.e. type of post). Thus, we define *perplexity* in this case to be a measure of *proximity* to our test set.

3 Methods

3.1 Data Source Overview

We wanted the data for our analysis to include a good representation of the online community and also have a lot of data to ensure that each SLM could be adequately trained. Reddit users fit our needs well as Reddit is an anonymous platform where people feel free to express themselves without censorship. There are no additional restrictions that would make us believe that user activity would behave differently from the users of interest in the (Danescu-Niculescu-Mizil et al, 2013) study. In order to capture discussions related to politics, we filtered the Reddit comment data to include only the r/politics subreddit. We wanted to capture discussion surrounding an entire election cycle, so we cast a wide net and included the election year, as well as the year prior to and the year following each election. We organized data by month. Table 1 shows the key statistics of our data for the 2016 Presidential election period from 2015 to 2017. Figure 2 shows the number of posts in the politics subreddit by month.

From the table we can see that over 40,000 users have posted more than 50 times. This gives us the ability to observe user-level characteristics. In addition, the Reddit data we have access to goes back to 2005 and also covers a variety of topics. This is very useful as it allows us to provide two different baselines to compare our SLM data to. We plan to implement SLMs for the following specific datasets:

- r/politics, January 2015 - December 2017
- r/politics, January 2011 - December 2013

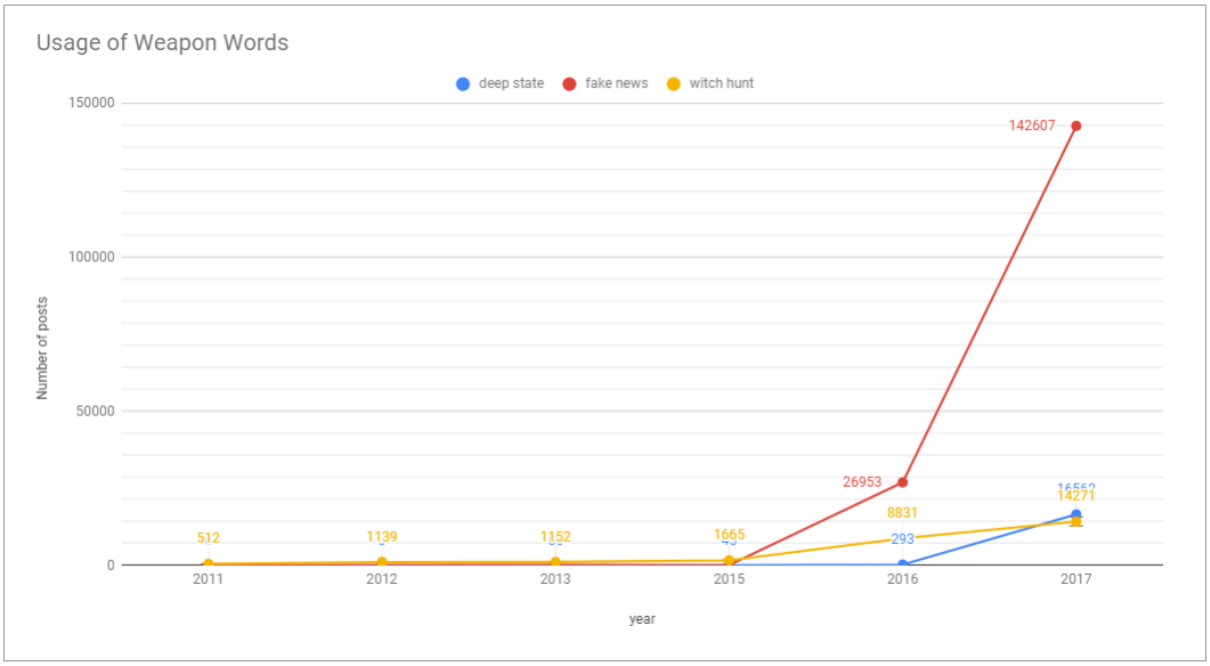

Figure 1: Number of Posts including Weapon Words, by year

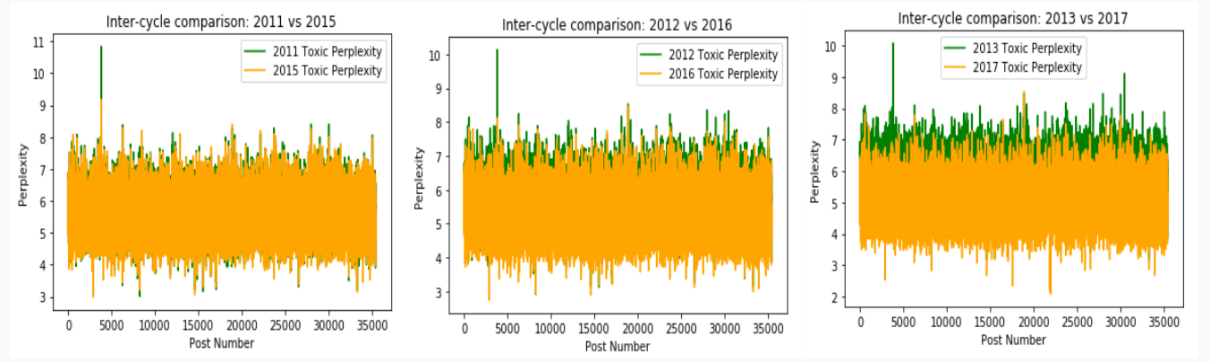

Figure 2: Perplexity scores of 2018 test set using SLMs, averaged by year for each post in test set. Each plot shows comparison of scores for two different years.

Total Number of Posts	44,883,364
Number of users	840,870
Users with more than 50 posts	41,066

Table 1: 2016 Presidential Election Reddit Data Summary

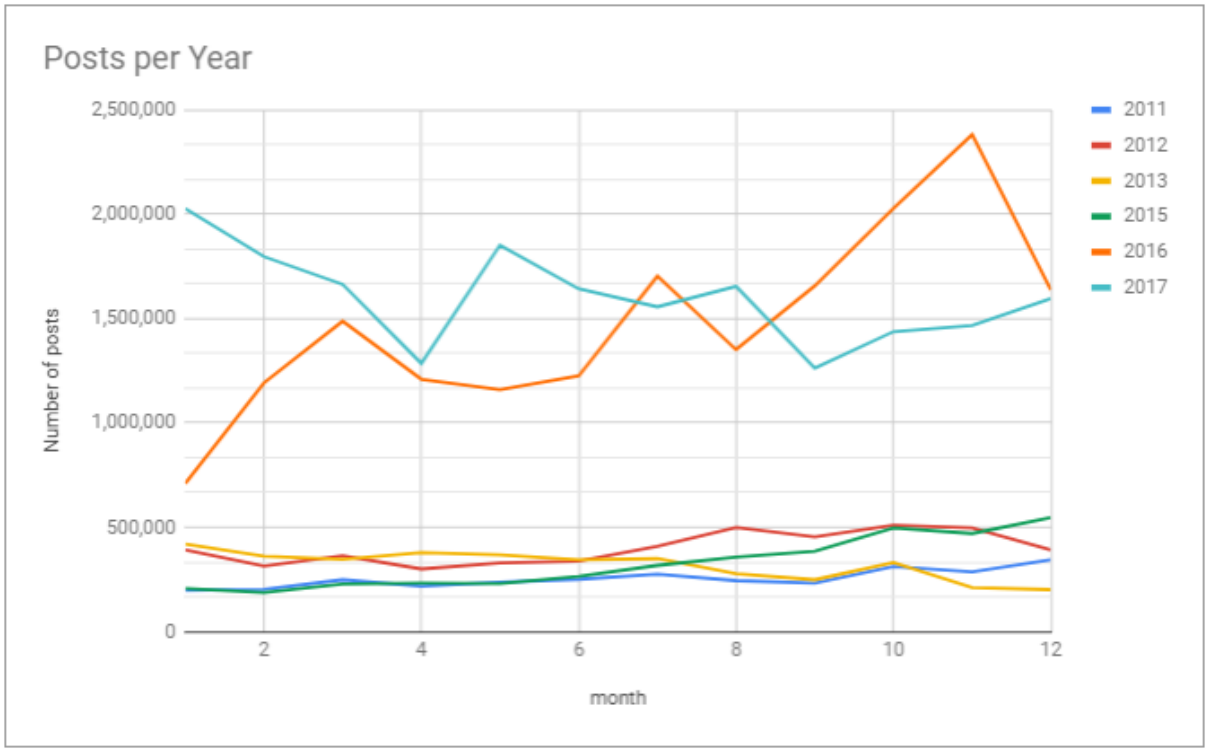

Figure 3: Total number of posts by month in the r/politics subreddit.

3.2 Weapon Words

In order to analyze linguistic change in relation to Trump Weapon Words, we have identified several key phrases,WW_i that have been coined or repeatedly used by Donald Trump during different time period of his Presidency and presidential campaign. The following weapon words were used, however, we could have easily expanded our collection of weapon words for this study:

- "fake news"
- "witch hunt"
- "deep state"

We generated a test set which consisted of posts made in the r/politics subreddit in 2018 which include any combination of one more of WW_i. Table 2 shows test for 10 random posts from the test set. We used the measure of perplexity to test whether or not posts containing these Weapon Word were being used more often than those without. From the test set, we generated a *Perplexity* score for each post for each SLM to see how closely the SLM resembled the language of the posts in the test set.

3.3 Community-Level Effects

For each $SLM_{m(p)}$, we calculated the average perplexity of posts, p_i. We compared these to a baseline which was the 2012 Election set of SLMs, testing the hypothesis that 2016 Presidential Election era posts would have a decreasing perplexity and smaller perplexity than the SLM's generated during the 2012 Presidential Election.

4 Results and Discussion

Our study shows that there has indeed been a shift in social communities' use of language towards that of the President. Using the test set of posts from 2018, we have tested the performance of SLMs for 2016 Election era months against performance of 2012 Election era months. Our expection based on the frequency of WW_i, was that the strongest differences would be observed in 2017 compared to 2012 Election models. We ran tests for several different time period and resolutions to best understand when the shift in linguistic change happened and how strong it was. In Figure 4, we can see the 2012 Election behaves as a constant. Picking back up in 2015, perplexity values are still very close to where they were during the 2012 Election. It is only a few months later in 2015 when we see a downward trend that continues

Author	Text
Benjanon_Franklin	No she just mishandled top secret info and covered the evidence up by deystroying...
older_than_dirt	OMG you people are actually off the rails insane. President Trump is not going to...
callahan09	It's funny because that unlimited data collection from people's social media usage is...
PresidentBiglyhands	Good. Give the Deep State the warning shot they so clearly need. They cross...
renew123	Serious question for the lefties here... with all the big crack down on fake news, ...
Alchemist2121	Nah that's too overt. more like a fully funded department of MS that focused on...
RainbowDarter	They know, but attribute that to defending from the corrupt left and deep state...
EvryMthrF_ngThrd	"O Country Where Art Thou?" *Original Motion Picture
fuzeebear	Rigged Russia Witch Hunt He keeps extending his clever little nickname for the ...
slutttt	Break out the popcorn, folks! Here's this weekend's tweet forecast: Saturday: Huge ...

Table 2: 10 posts from test set: posts from 2018 containing weapon words.

through the 2016 Election Day in November and then plateaus. This distinction is also presented in Figures 5 and 6 with a density histogram for each of the 3 years for each election cycle. The 2012 Election cycle years exhibit virtually no change, while the 2016 Election cycle shows a downward shift from year to year.

Tables 4 and 5 contain text for the 5 highest and lowest scoring posts for SLM_{Nov17}.

4.1 Natural Linguistic Change

The original study conducted by (Danescu-Niculescu-Mizil et al, 2013) revealed a natural evolution of language used in online discussions. For our results, we did not attempt to control for natural linguistic changes in each SLM and how any such changes within the test set posts might affect the results. It's likely that there is a small bias at work in favor of lower *Perplexity* scores in more recent SLMs and the 2016 Election era collection overall. We can, however, be certain that the proximity of online political discussions has made a significant shift towards the language identified by (Lakoff, 2016) as *weapon words*.

5 Conclusion

Lakoff's insights were based on intuition and anecdotal evidence, and he is right. Two things have happened to online discussions that strongly support his hypothesis. One, the frequency at which online communities discuss politics has risen substantially since Trump began weaponizing his words beginning with his presidential campaign in 2015. Second, the *weapon words* used by the President have become adopted by online communities, and linguistic change has shifted towards his speech patterns. We see potential for many other use cases and opportunities to continue analysis using SLMs to determine with more precision when online communities are adopting language. In addition, the presense of bots has clouded our ability to understand true user behavior. We propose that continued study into this space include the detection of automated content and its role in developing linguistic change patterns. Some of the 2016 and 2017 SLMs indicated that individual posts with the highest proximity (or lowest perplexity) were generated by bots (see Table 5). We believe that combining SLM techniques such as we've used with state of the art NLP bot detection methods developed by (Kumar et al., 2017) would be a powerful step forward in gaining more understanding of our language patterns in modern communication mediums.

6 Source Code

All project source code and materials are available at: Linguistic Change in Online Political Discourse

References

Cristian Danescu-Niculescu-Mizil et al. 2013. *No Country for Old Members: User Lifecycle and Linguistic Change in Online Communities*.

Srijan Kumar et al. 2017. *An Army of Me: Sockpuppets in Online Discussion Communities*.

Stanley Chen and Joshua Goodman. 1998. *An Empirical Study of Smoothing Techniques for Language Modeling*. Center for Research in Computing Technology, Harvard University.

George Lakoff. 2016. *How You Help Trump*.

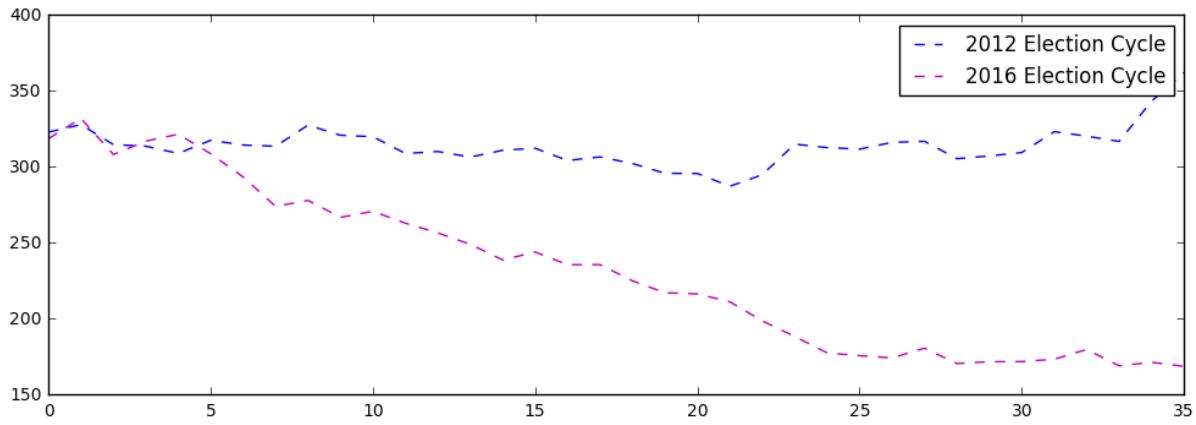

Figure 4: Perplexity scores for each SLM (Y axis = Perplexity, X axis = month in election cycle)

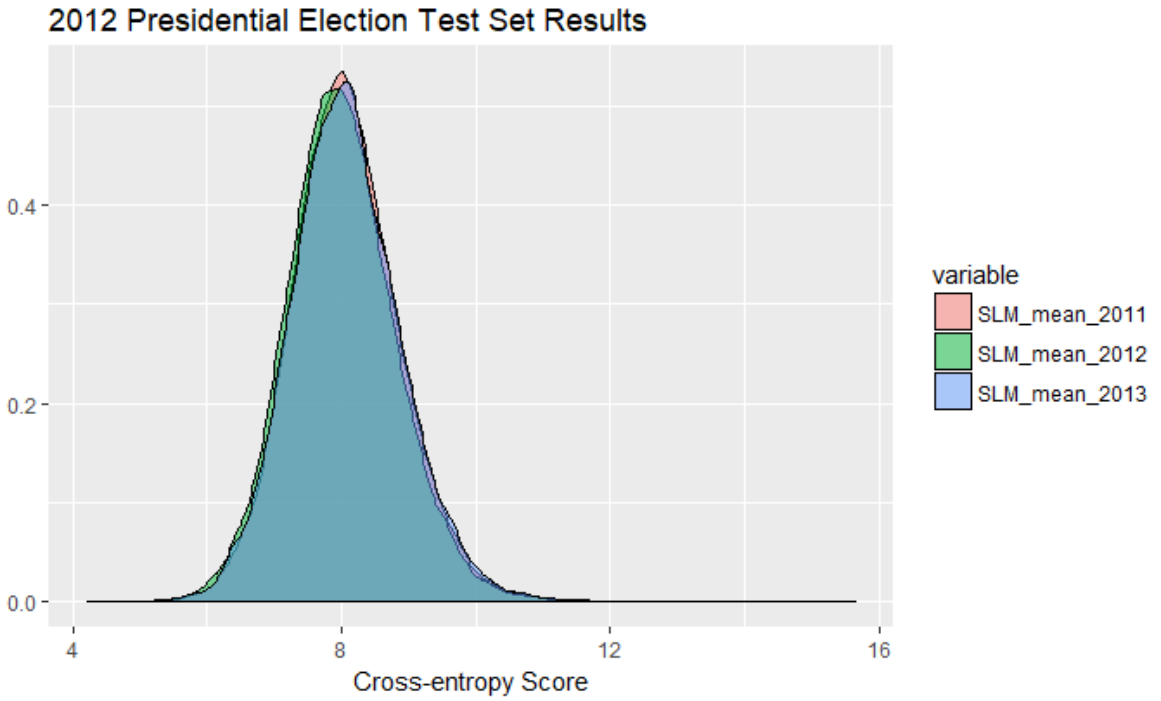

Figure 5: Density plot of histogram for each year surrounding 2012 Presidential Election

Author	Text
AutoModerator	Hi ‘RejectForce‘, your post ‘This is what fake news looks like‘ has been rem...
AutoModerator	Hi ‘AimingWineSnailz‘, your post ‘Fake News, Part 1: Origins and evolution‘ ...
AutoModerator	Hi ‘Cotton9‘, your post ‘Dr. Pieczenik: Trump Has Officially Destroyed the D...
AutoModerator	Hi ‘timekill05‘, your post ‘Trump Walks Out On 1990 CNN Interview For Being ...
AutoModerator	Hi ‘goose7771‘, your post ‘Sean Hannity finds out on-camera that Trump tryin...

Table 3: Posts with Top 5 Perplexity Scores from November 2017 SLM

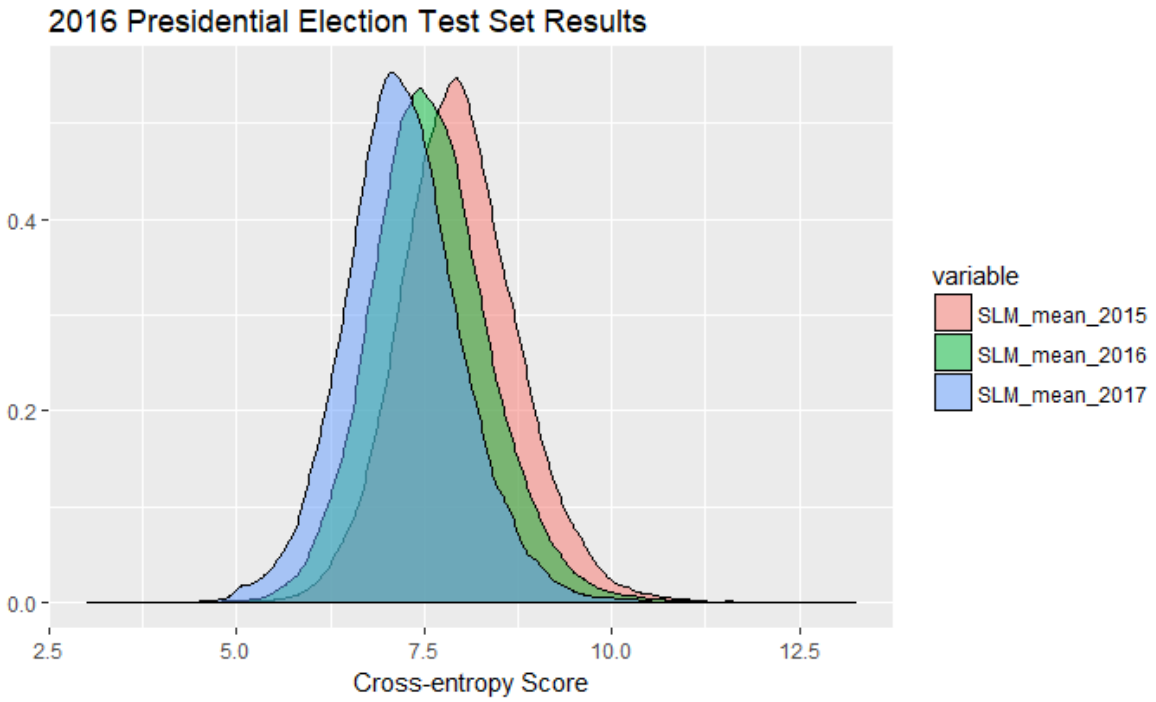

Figure 6: Density plot of histogram for each year surrounding 2016 Presidential Election

Author	Text
Apostate1123	Bob Mueller nick name odds Liddle Bob Mueller 2-1 Russia Hoax Robert 3...
Qweef	*sneakily* Dudes distracting us at every chance he can get RIGHT TO CH...
MrMadcap	The important thing to always remember is that: Hillary. Obama. Mueller. ...
TDisacuckfactory	LOL DAE GEORGE SOROS CRISIS ACTOR DEEP STATE PLOT FLAT ...
MrsMI1UCAN2	Shh. They speak of *her*, our true deep state overlord, let us recognize ...

Table 4: Posts with Lowest 5 Perplexity Scores from November 2017 SLM