Dynamic Topic Modelling: A Case Study on Russian Twitter Troll Data

Chunxia Cao, Bo Chai, Nicholas Chao, Johnny Chen, Liangkai Hu, Shaobo Liang, and Stephen Newns
Georgia Institute of Technology - CSE 6242 Spring 2019

Introduction

Misinformation has become a strategic tool used by foreign countries to interfere with elections. Social media platforms such as Facebook and Twitter allows individuals to quickly spread false information. The effects can be damaging as seen in the U.S. 2016 Presidential election, where partisan divisions were facilitated by Russian “trolls”. There is a need for effective tools to prevent foreign entities from exploiting social media platforms for their own agenda. Topic modelling is a widely used method for studying the content of messages. The analysis, however, is often static and does not provide a temporal view of the problem. In this project we develop a dynamic topic modelling tool to allow users to explore and analyze large social media data to better understand the topics that internet trolls use in their misinformation campaigns. We apply it to an open source dataset from Twitter.

Problem Definition

We are developing a visual tool to analyze the Russian Twitter troll data released by FiveThirtyEight/Clemson University, implicating the Internet Research Agency of interfering with the 2016 US Presidential election. This tool will allow users to uncover topics and sentiment found in these tweets and their evolution over time. Additionally, we will explore how discussion varies amongst different groups and gleam information that can be utilized to identify the impact of major news events.

Survey

In the past decade, participation in online social networks has exploded. The study of social behaviors on social media has become a significant area of academic interest. Previous research has focused machine learning for network analysis (Lancichinetti and Fortunato, 2009). Galán-García et al., (2016) applied a supervised classification approach to detect Twitter cyberbullies. Our case study is the Russian Twitter troll accounts that potentially influenced the 2016 US Presidential election, on which multiple studies have recently been published. Griffin and Bickel (2018) utilized unsupervised machine learning to extract relevant information about them, Linvill and Warren (2018) employed classification methods to identify five categories of accounts, Llewellyn et al. (2018) found that behavior of tweets radically changed during the day of the referendum to amplify content produced by trolls, and Badawy et al. (2018) applied label propagation and geo-spatial analysis on where retweets were coming from and to classify users as liberal or conservative with high precision and recall. However, prior analyses only span a few weeks, so long-term temporal analysis has not been implemented. With tweets spanning the entire election, we can analyze long-term trends.

Considering sentiment analysis, Cheong and Cheong (2011) revealed benefits of active Twitter involvement during emergency situations, Smailović et al. (2014) studied the evolution of tweet sentiment on selected companies to forecast stock prices using active learning strategies, which could be further leveraged for temporal topic analysis. Kant et al. (2018) developed a flexible and practical sentiment analysis approach using Plutchik's wheel of emotion as a framework for tagging tweets. This pre-trained model, particularly well-suited emotion-classification, could be useful for this project. Also, studies were done on time-sensitive similarity and sentiment analysis on bots in networks (Bessi and Ferrara, 2016; Sakaki et al., 2013) and on detection and categorization of bot influence (Kim et al. (2019)), which gleaned insights on election-related behaviors in networks.

Stewart et al. (2018) were able to identify Russian trolls' central positions in retweet networks. Automated authorship detection is a powerful tool for network analysis (ACBIT, 2019), in which text mining and supervised learning have been applied to investigate various network analysis to explore real authors behind text (Chollet, 2018; Savage et al., 2017). Im et al. (2019), made attempts to identify troll accounts through regression & decision tree classification and clustering learning, achieving 78.5%

precision and 98.9% AUC out-of-sample, utilizing, five features, while Tärning (2017) emphasized source usages and achieved mixed results. Additionally, methods could be extended by incorporating deeper linguistic semantic features (Gómez-Adorno et al., 2016).

Dynamic analysis and network evolving topology studies are also helpful to network analysis. Albert and Barabasi (2000) proposed a continuum theory to predict how networks grow and evolve and the scaling function to show, depending frequency of these processes, two topologically-different networks can emerge and their connectivity distribution. Blei and Lafferty (2007) proposed a dynamic topic modeling approach based on the Latent Dirichlet Allocation (LDA) algorithm that studies how words most associated with a given topic to evolve with time. Due to the short length, typos, and colloquialisms in tweets, extra data cleaning and fewer tokens would be required. Ritterman et al. (2009) presented an approach to predict some dynamic variables of swine flu using volume of tweets, wherein SVMs were effective to extract information from high-dimensional, sparse data, which could be applied in our project.

Proposed Methods

Intuition:

While this dataset is fairly new, it is already becoming a focus of academic research, as evident by our literature survey. From what we have found, little work has been done to explore evolution of topics and sentiment present in tweets from a temporal point of view. There has been non-temporal topic analyses (Linvill & Warren, 2018) (Tarning, 2017), while most temporal analyses were related to pure word frequency (Llewellyn et al, 2019). Kim et al, 2019, analyzed tweet contents temporally through the lens of Active-Network Theory, uncovering strategic behavior between troll accounts, but we aim to provide a more general perspective of major topics found in the data. Developing a visually-communicative tool will be an effective way of exploring these tweets from a perspective not yet seen. Our approach provides the following innovations:

1) Allows for temporal analysis by providing a visualization that dynamically updates over time.
2) The user can explore topics at specific time points for correlation with major news events.
3) The content is displayed to allow users to identify patterns and groups of content in an intuitive way.

Tableau and R are used for the bulk of the analysis and visualization of the dataset.

Methodology:

In order to model the temporal effects on topics we use Structural Topic Models (STMs), a variant of LDA that incorporates covariate effects from document-level metadata on each document's dirichlet prior in terms of topic-content (Roberts et al., 2014). Robert et al. proposed the algorithm as an effective way to analyze open-text survey response data in social-science research. We used the publish month of the tweet as a document-level covariate. In doing so, we can see how a document's dirichlet prior distribution of topics and words is impacted by the publish month of the tweet. In doing so, we can explore which words had a high covariate effect for that model in that particular month. We used the *stm* package to carry out this analysis, an open-source implementation of STMs in R.

Given the large amount of documents we are working with, determining the right number of topics in the model is a non-trivial task. A standard way of selecting number of topics is to train several models with different number of topics and compare the models using metrics such as semantic-coherence and exclusivity. This would not be a practical approach as effective topic modeling is an iterative process, requiring one to build topic models to learn more about the data, and adjust the data to account for issues such as including additional stop-words and removing outlier documents that drastically bias the model. As model training time surpasses several hours, a bottleneck is created in model development.

In order to circumvent this bottleneck, we utilized a heuristic proposed by Mimno and Lee that determines an appropriate number of topics through a t-SNE, t-distributed neighbor embedding, of the word co-occurrence matrix, in which anchor-words are found (Mimno and Lee, 2014). The number of anchor-words is the number of topics selected, and these anchor-words are used as a starting point in the expectation-maximization optimization of the structural topic model. This process still look 8.2 hours to train on the data, however the alternative of searching through many values of K would likely have taken several days. This heuristic is available through the *stm* and *Rtsne* packages in R.

We used only English tweets for our analysis, and tweets from January 1st, 2016 through January 1st, 2017. We also found accounts would often spam the same tweet repeatedly in a single day, which would introduce an unwanted bias in the topic models. When this occurred, we only kept the first tweet. Additionally, we removed retweets and tweets that only contained links to other webpages, as proved difficult to incorporate in the topic models effectively. After doing this, we were left with a sample of 182,327 tweets.

When tokenizing the documents we extracted uni-grams, bi-grams, and tri-grams after converting words to their lemmas. We also converted common internet colloquialisms to their proper english versions (e.g. “lol” to “laugh out loud”. After taking these steps and cleaning the data of non-numeric and non-character strings filtered out documents that occur less than 15 documents. After taking these steps were left with 12,942 tokens, giving us a model dataset that is a 182,327 x 12,942 document-feature matrix.

Visualization methods

To understand the twitter data, exploratory analysis was performed and summarized in three Tableau dashboards[1]. The twitter accounts were categorized by political leanings (left vs right) by Clemson University researchers. The categories are displayed in this dashboard to visualize the distribution of different categories over the time period of a year. The bar plots show that a few users have the majority of the followers. We also include and option to filter the visualizations by the Twitter account categories of “Right Troll” versus “Left Troll” in order to observe how the strategies of different types of trolls evolve in response to different topics and to major news events.

For detailed analysis of specific topics, we developed an interactive dashboard to explore the topics over time windows[2] [3]. The user can perform focused analysis on a specific topics from a list of topics discovered via our analysis. For example, the user can choose “Election News”, and visualize how the topic prevalence/number of tweets changed over time. On a more granular level, the user can also click on “Election News”, then choose any relevant key terms prevalent in this topic, and visualize the terms’ rankings over time. Additional analysis such as a snapshot of top seven ranked terms in a topic at a chosen month is also available on the dashboard. We experimented with the most effective visualization of the top terms within each topic, and chose a word cloud that changes with respect to the time window and topic selected. A scrollbar is available to visualize the change in key topics by scrolling through time, enabling cross-sectional time series analysis.

The final dashboards are available to the public on Tableau Public.

Experiments and Evaluations

Our project will help users understand how topics present in relevant tweets change over time in terms of their content. In this pursuit, we can determine success if the model can identify

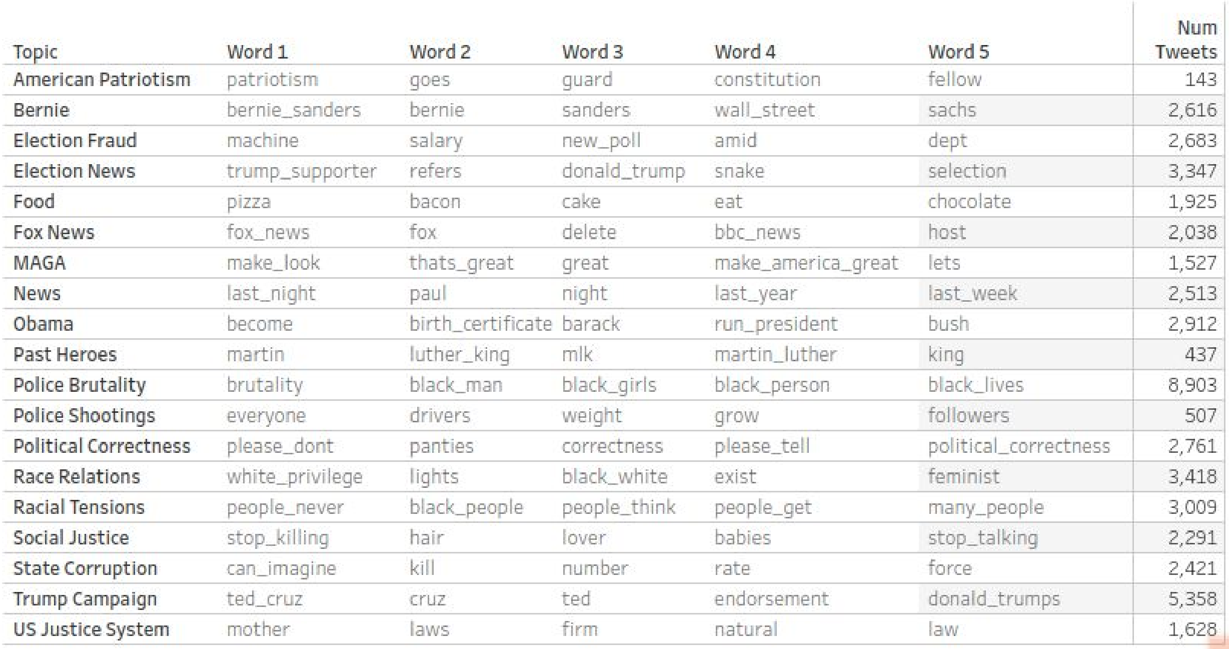
[1] https://public.tableau.com/profile/shaobo7549#!/vizhome/EDA_15557961312920/EDA?publish=yes

[2] https://public.tableau.com/profile/johnnychen#!/vizhome/troll_topicsv2/Dashboard1?publish=yes

[3] https://public.tableau.com/profile/schai5500#!/vizhome/word_importance_temporal_analysis_updated/WordImportance-Dashboard?publish=yes

non-random effects on topical content from the publish month of a tweet, and evaluate if those changes are coherent through human judgement.

Our final model contains 67 topics. While several topics seem in-coherent, many were obviously distinct. Figure 1 displays the most probable words to be drawn from the most human-interpretable topics. While it was sometimes difficult to interpret the effect of the time covariate on topic content, however we were able to find some clearly meaningful effects occurring. For instance, in the topic of *Obama* we found that one of the strongest covariate effects for November of 2016 was from the unigram “former” (figure 2). This corresponds to Trump’s victory in the US Presidential election that month, marking the end of the Obama era. Another example, also related to Trump’s election was the strong covariate effects observed on the US Justice System topic, related to James Comey. This indicates that during election time discussions around the US justice system were centered on FBI Directory, James Comey.



Topic	Word 1	Word 2	Word 3	Word 4	Word 5	Num Tweets
American Patriotism	patriotism	goes	guard	constitution	fellow	143
Bernie	bernie_sanders	bernie	sanders	wall_street	sachs	2,616
Election Fraud	machine	salary	new_poll	amid	dept	2,683
Election News	trump_supporter	refers	donald_trump	snake	selection	3,347
Food	pizza	bacon	cake	eat	chocolate	1,925
Fox News	fox_news	fox	delete	bbc_news	host	2,038
MAGA	make_look	thats_great	great	make_america_great	lets	1,527
News	last_night	paul	night	last_year	last_week	2,513
Obama	become	birth_certificate	barack	run_president	bush	2,912
Past Heroes	martin	luther_king	mlk	martin_luther	king	437
Police Brutality	brutality	black_man	black_girls	black_person	black_lives	8,903
Police Shootings	everyone	drivers	weight	grow	followers	507
Political Correctness	please_dont	panties	correctness	please_tell	political_correctness	2,761
Race Relations	white_privilege	lights	black_white	exist	feminist	3,418
Racial Tensions	people_never	black_people	people_think	people_get	many_people	3,009
Social Justice	stop_killing	hair	lover	babies	stop_talking	2,291
State Corruption	can_imagine	kill	number	rate	force	2,421
Trump Campaign	ted_cruz	cruz	ted	endorsement	donald_trumps	5,358
US Justice System	mother	laws	firm	natural	law	1,628

Figure 1: We found several clearly coherent topics in our STM

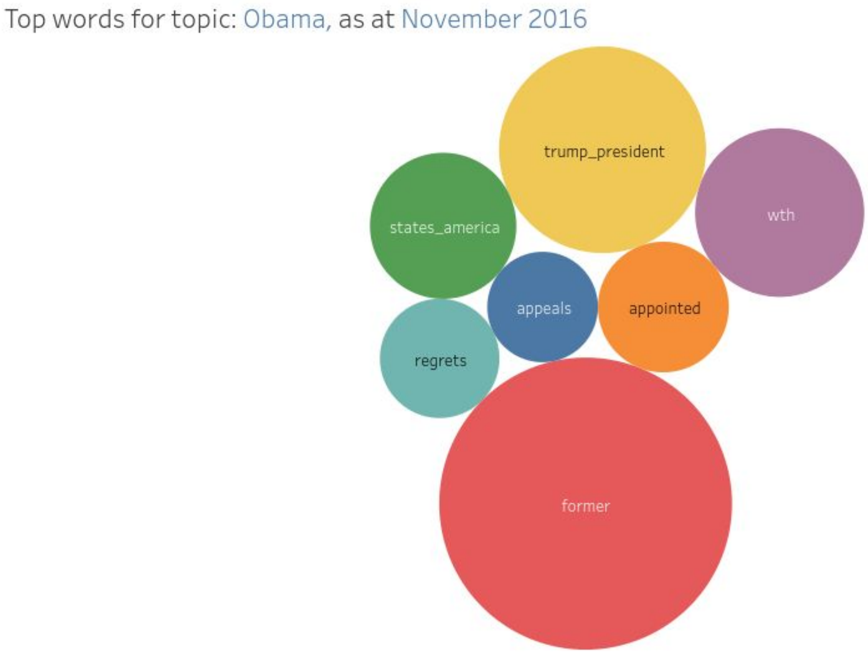


Figure 2:

‘former’ has a strong covariate effect on the Obama topic, corresponding to Trump’s victory signaling the end of the Obama era

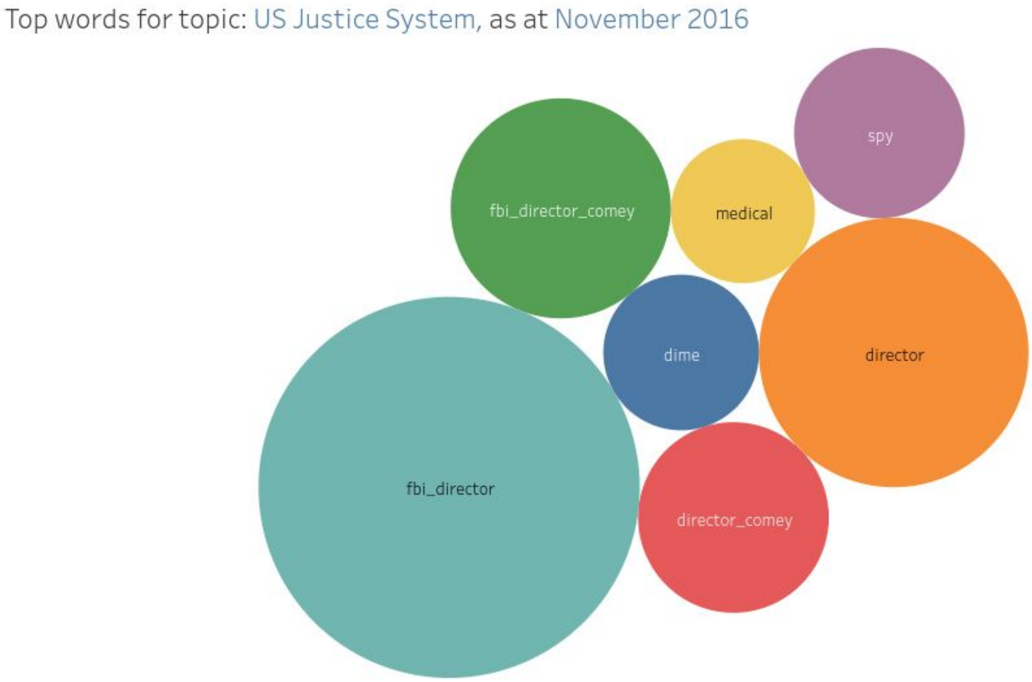


Figure 3: ‘fbi_director’ and ‘director_comey’ have strong covariate effects on the US Justice System topic at the time of Trump’s Election

Conclusions and discussion

We successfully incorporated temporal dynamics of topic content into our topic model and the Tableau tool we built effectively allows users to explore those dynamic effects. While the dynamic effects were often difficult to comprehend, some meaningful effects were found, such as the high probability of the word “UK” in the month of June for the topic of Elections. This effect is likely caused by the Brexit referendum that occured that month. We have found that a word-cloud representing the most probable words to be drawn from a topic, given the time period selected, is an effective way of exploring these topic models.

Further analysis should explore the possibility of incorporating time into the topic and word distributions in a way that the topic and word distributions are dependent on those of the previous time frame. This approach simply considered the publish month as a categorical factor, and so adjacencies between time slices are not taken into account in the topic prevalence and content. These adjacencies might be important to capture if topic prevalence or content tends to grow in a sequentially dynamic way.

References

1) ACBIT (2019). Authorship Attribution with Python. Available at: http://www.aicbt.com/authorship-attribution/

2) 2)Albert, R., & Barabási, A. L. (2000). Topology of evolving networks: local events and universality. Physical review letters, 85(24), 5234.

3) Badawy, A., Ferrara, E., & Lerman, K. (2018, August). Analyzing the digital traces of political manipulation: The 2016 russian interference twitter campaign. In 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp. 258-265). IEEE.

4) Bessi, A., & Ferrara, E. (2016). Social bots distort the 2016 US Presidential election online discussion.

5) Blei, D. M., & Lafferty, J. D. (2006, June). Dynamic topic models. In Proceedings of the 23rd international conference on Machine learning (pp. 113-120). ACM.

6) Blei, D. M., Ng, A. Y., Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research 3, 993-1022.

6) Cheong F., Cheong C. (2011). Social media data mining: a social network analysis of tweets during the Australian 2010–2011 floods. In: 15th Pacific Asia conference on information systems (PACIS). Queensland University of Technology, pp 1–16.

7) Chollet F. (2018). Deep Learning with R, Chapter 3. Manning Publications.

8) Galán-García, P., Puerta, J. G. D. L., Gómez, C. L., Santos, I., & Bringas, P. G. (2016). Supervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying. Logic Journal of the IGPL, 24(1), 42-53.

9) Gómez-Adorno, H., Sidorov, G., Pinto, D., Vilariño, D., & Gelbukh, A. (2016). Automatic authorship detection using textual patterns extracted from integrated syntactic graphs. Sensors, 16(9), 1374.

10) Griffin, C., & Bickel, B. (2018). Unsupervised Machine Learning of Open Source Russian Twitter Data Reveals Global Scope and Operational Characteristics. arXiv preprint arXiv:1810.01466.

11) Im, J., Chandrasekharan, E., Sargent, J., Lighthammer, P., Denby, T., Bhargava, A., ... & Gilbert, E. (2019). Still out there: Modeling and Identifying Russian Troll Accounts on Twitter. arXiv preprint arXiv:1901.11162.

12) Kant, N., Puri, R., Yakovenko, N., & Catanzaro, B. (2018). Practical Text Classification With Large Pre-Trained Language Models. arXiv preprint arXiv:1812.01207.

13) Kim, D., Graham, T., Wan, Z., & Rizoiu, M. A. (2019). Tracking the Digital Traces of Russian Trolls: Distinguishing the Roles and Strategy of Trolls On Twitter. arXiv preprint arXiv:1901.05228.

14) Lancichinetti, A., & Fortunato, S. (2009). Community detection algorithms: a comparative analysis. Physical review E, 80(5), 056117.

15) Linvill, D. L., & Warren, P. L. (2018). Troll Factories: The Internet Research Agency and State-Sponsored Agenda Building. pwarren. people. clemson. edu/Linvill_Warren_TrollFactory. pdf.

16) Llewellyn, C., Cram, L., Favero, A., & Hill, R. L. (2018). For whom the bell trolls: Troll behaviour in the Twitter Brexit debate. arXiv preprint arXiv:1801.08754.

17) Mimno, D., & Lee, M. (2014), Low-dimensional Embeddings for Interpretable Anchor-based Topic Inference. Proceedings of the 2014 Conference on EMNLP, 1319-1328.

17) Ritterman, J., Osborne, M., & Klein, E. (2009, November). Using prediction markets and Twitter to predict a swine flu pandemic. In 1st international workshop on mining social media (Vol. 9, pp. 9-17). ac. uk/miles/papers/swine09. pdf (accessed 26 August 2015).

18) Roberts, M., Stewart, B., & Tingley, D., (2014). stm: R Package for Structural Topic Models. Journal of Statistical Software.

19) Sakaki, T., Okazaki, M., & Matsuo, Y. (2013). Tweet analysis for real-time event detection and earthquake reporting system development. IEEE Transactions on Knowledge and Data Engineering, 25(4), 919-931.

20) Savage, D., Wang, Q., Zhang, X., Chou, P., & Yu, X. (2017, March). Detection of Money Laundering Groups: Supervised Learning on Small Networks. In Workshops at the Thirty-First AAAI Conference on Artificial Intelligence.

21) Smailović, J., Grčar, M., Lavrač, N., & Žnidaršič, M. (2014). Stream-based active learning for sentiment analysis in the financial domain. Information sciences, 285, 181-203.

22) Stewart, L. G., Arif, A., & Starbird, K. (2018, February). Examining trolls and polarization with a retweet network. In Proc. ACM WSDM, Workshop on Misinformation and Misbehavior Mining on the Web.

23) Tärning, J. (2017). Troll Detection: A study of source usage between clusters of Twitter tweets to detect Internet trolls.