



I N S T I T U T E F O R D E F E N S E A N A L Y S E S

Unsupervised Machine Learning Reveals How the Twitter Topic Patterns of the Russian-Backed Internet Research Agency (IRA) Evolved Over Time (Presentation)

Emily M. Parrish
Shelley M. Cazares
Jenny R. Holzer

January 2020

Approved for public release;
distribution is unlimited.

IDA Document NS D-12056

Log: H 20-000024

INSTITUTE FOR DEFENSE ANALYSES
4850 Mark Center Drive
Alexandria, Virginia 22311-1882



The Institute for Defense Analyses is a nonprofit corporation that operates three Federally Funded Research and Development Centers. Its mission is to answer the most challenging U.S. security and science policy questions with objective analysis, leveraging extraordinary scientific, technical, and analytic expertise.

About This Publication

This work was conducted by the Institute for Defense Analyses Central Research Program. The views, opinions, and findings should not be construed as representing the official position of either the Department of Defense or the sponsoring organization.

For More Information

Shelley M. Cazares, Project Leader
scazares@ida.org, 703-845-6792

Leonard J. Buckley, Director, Science and Technology Division
lbuckley@ida.org, 703-578-2800

Copyright Notice

© 2020 Institute for Defense Analyses
4850 Mark Center Drive, Alexandria, Virginia 22311-1882 • (703) 845-2000.

This material may be reproduced by or for the U.S. Government pursuant to the copyright license under the clause at DFARS 252.227-7013 (Feb. 2014).

Executive Summary

BACKGROUND: Information Operations (IO) are a key component of our adversaries' strategy to undermine U.S. military power without escalating to more traditional (and more easily identifiable) military strikes. Social media activity is one method of IO. In 2017 and 2018, Twitter suspended thousands of accounts likely belonging to the Kremlin-backed Internet Research Agency (IRA). Clemson University archived a large subset of these tweets (2.9M tweets posted by over 2800 IRA accounts), tagged each tweet with metadata (date, time, language, supposed geographical region, number of followers, etc.), and published this dataset on the polling aggregation website *FiveThirtyEight*.

METHODS: Machine Learning researchers at the Institute for Defense Analyses (IDA) downloaded Clemson's dataset from *FiveThirtyEight* and analyzed both the content of the IRA tweets and their accompanying metadata. Using unsupervised learning techniques (Latent Dirichlet Allocation), IDA researchers mapped out how the patterns in the IRA's tweet topics evolved over time.

RESULTS: Results showed that the IRA started tweeting in/before February 2012, but ramped up significantly in May/June 2015. Most tweets were in English, and most likely targeted the U.S. The IRA created new accounts after the first Twitter suspension in November 2017, with each new account quickly establishing an audience. Between at least January 2015 and October 2017, the IRA's English tweet topics evolved over time, becoming tighter, more specific, more negative, and more polarizing, with the final pattern emerging in late 2015.

DISCUSSION: The United States government must expect that our adversaries' social media activity will continue to evolve over time. Efficient processing pipelines are needed for semi-automated analyses of time-evolving social media activity.

Unsupervised Machine Learning Reveals how the Twitter Topic Patterns of the Russian-backed Internet Research Agency (IRA) Evolved over Time

Emily Parrish, Shelley Cazares, Jenny Holzer
Institute for Defense Analyses (IDA)

22 Jan 2019

The Institute of Defense Analyses (IDA) is a non-profit corporation operating three Federally Funded Research and Development Centers (FFRDCs). IDA's mission is to provide objective analyses of national security issues and related challenges, particularly those requiring scientific, technical, and analytic expertise.

In Fiscal Year 2019, IDA devoted limited internal research and development funds to exploring Russian-backed activity on Twitter against the United States. IDA used a data-driven, quantitative approach, employing unsupervised machine learning techniques and other data analytics methods. IDA focused on the *evolution* of the Russian-backed activity, extracting and tracking how topic patterns in that activity *changed over time*. These slides summarize our analysis.

We thank the following internal IDA reviewers for their suggestions and guidance:

- Richard Anderson
- Claudio Biltoc
- Arun Maiya
- Prashant Patel

- Propaganda happens.



German **leaflet**
dropped on
British soldiers
during WWII

Propaganda happens.

It's not new.

This is a leaflet¹ dropped by the German military onto British warfighters during World War II. It states, “British soldiers! You are fighting and dying far away from your country while the Yanks are putting up their tents in Merry Old England. They’ve got lots of money and loads of time to chase after your women.” Presumably, this message was an Axis attempt to sow discord between British and American allies.

¹ Image number 137109, NAM 2002-02-916-1, National Army Museum, London UK. <https://collection.nam.ac.uk/detail.php?acc=2002-02-916-1>

- Propaganda happens.



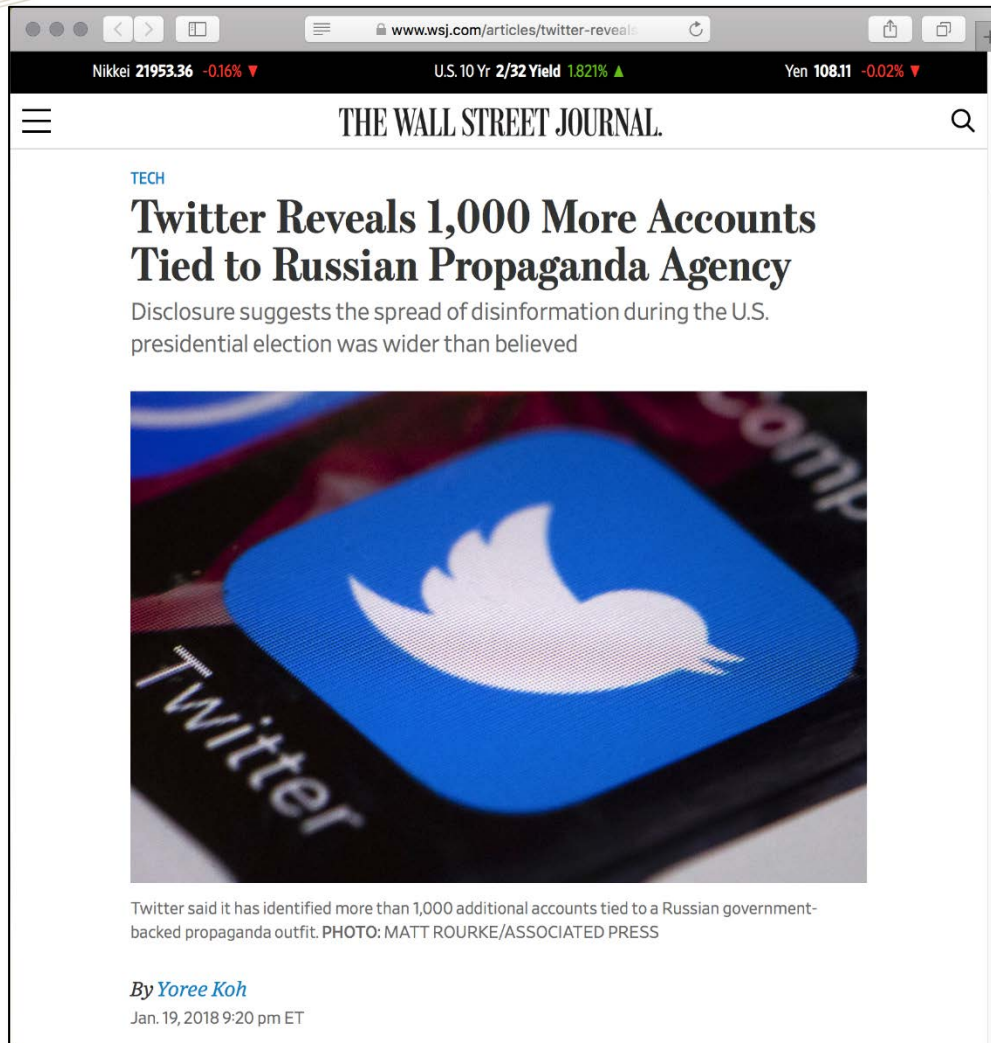
American-backed
radio station
in Eastern Europe
during the Cold War

Other nations have also used propaganda for national security aims, including the United States.

This poster² advertises Radio Free America, an American-backed radio station that gained prominence in Eastern Europe during the Cold War.

² Velkey, Robert. 2017 Oct 06. Today in History: Radio Free Europe's Hungarian-Language Broadcast Began 66 Years Ago. *Hungary Today*.
<https://hungarytoday.hu/today-in-history-radio-free-europes-hungarian-language-broadcast-began-66-years-ago-20633/>

- **Propaganda happens.**

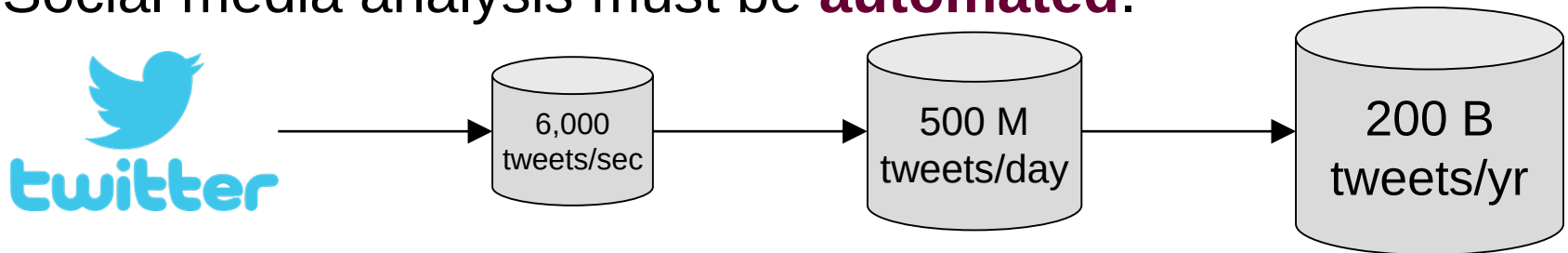


Russian-backed
social media
around the world
in the 2010s

More recently, news reports³ have discussed the Russian-backed use of social media around the world in the 2010s, in particular the Russian-backed *Internet Research Agency (IRA)*.

³ Koh, Yoree. 2018 January 19. Twitter Reveals 1,000 More Accounts Tied to Russian Propaganda Agency. *The Wall Street Journal*.
<https://www.wsj.com/articles/twitter-reveals-1-000-more-accounts-tied-to-Russian-propaganda-agency-1516414856>

- **Propaganda happens.**
- IRA Twitter tactics **evolved over time**:
 - Started tweeting in/before Feb 2012, but ramped up in May/Jun 2015
 - Most tweets in English, most likely targeted the U.S.
 - Created new accounts after first Twitter suspension in Nov 2017, each quickly establishing an audience
 - Between Jan 2015 – Oct 2017, English tweet topics evolved over time, becoming tighter, more specific, more negative, and more polarizing, with the final pattern emerging in late 2015
- Social media analysis must be **automated**.



- Efficient **processing pipelines** are needed for automated social media analysis → IDA has invested internal funds

As we will discuss, our analysis showed that IRA Twitter tactics evolved over time:

- The IRA started tweeting as early as February 2012, and then ramped up significantly in May/June 2015.
- Most IRA tweets were in English and most likely targeted the United States.
- The IRA quickly created new accounts after their first suspension from Twitter in November 2017, with each new account rapidly establishing an audience.
- Between at least January 2015 and October 2017, the IRA's English tweet topics evolved over time, becoming tighter, more specific, more negative, and more polarizing, with the final pattern emerging in late 2015.

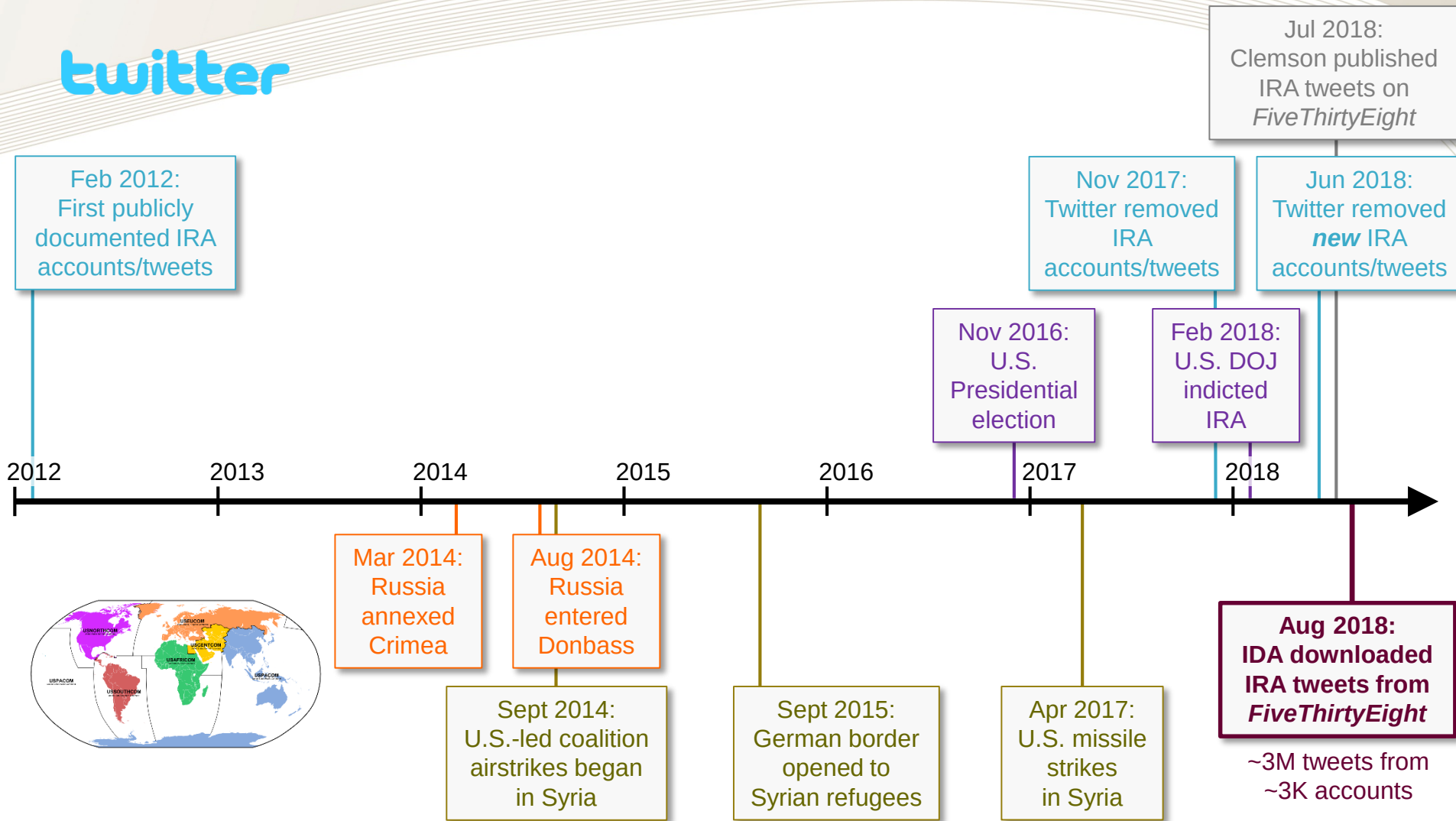
We used unsupervised machine learning and other data analytics techniques to arrive at these conclusions. Our techniques were semi-automated. Social media analysis techniques must be as automated as possible, due to the sheer size of the relevant data.

Considering Twitter as a whole (not just the IRA): In 2018, approximately 6,000 tweets were posted per second— that's 350 thousand tweets per minute, 500 million tweets per day, and 200 billion tweets per year.⁴ There simply aren't enough humans in the entire Department of Defense (DoD) and related organizations to manually assess that many tweets.

Therefore, an efficient processing pipeline is needed for automated or at least semi-automated analysis of social media. IDA has invested internal research and development funds to create such a pipeline.

⁴ Twitter Usage Statistics. *Internet Live Stats*. <https://www.internetlivestats.com/twitter-statistics/>

twitter



The Russian-backed IRA has been tweeting since (at least) 2012.
The U.S. took action after the 2016 Presidential election.

Before discussing our methods and results in detail, we first describe the specific set of tweets that we analyzed.

This slide shows a timeline of events, from 2012 through 2019. Below the timeline, we note a few events of DoD interest that took place in the real world.⁵ Above the timeline, we note events that took place on or in regards to Twitter:⁶

- (Blue, Left) The first IRA tweets in this dataset were posted in February 2012. The U.S. Senate Select Committee on Intelligence (SSCI) has stated that the IRA first focused on domestic Russian issues and audiences, turning later to foreign issues and audiences, such as in Ukraine.⁷
- (Orange) Events in Ukraine took place in 2014. Russia annexed the Crimea in March 2014⁸ and entered the Donbass in August 2014.⁹
- (Yellow) Events regarding Syria took place, as well. For example, American-led coalition airstrikes began in Syria in September 2014,¹⁰ the German border was opened to Syrian refugees in August/September 2015,¹¹ and American missile strikes took place in Syria in April 2017.¹²

Of course, many other real-world events of DoD interest also took place in this time frame— too many to note here. Through it all, the IRA was tweeting.

- (Purple) The United States did not take much action regarding the IRA's social media activities until fairly late in the game— namely, after the American Presidential election in November 2016. In February 2018, the U.S. Department of Justice (DoJ) indicted the IRA for conspiracy to defraud the United States and conspiracy to commit wire and bank fraud, in relation to the IRA's social media activities.^{13,14}

- (Blue, Right) Twitter itself began to take action on the IRA, as well. In November 2017, Twitter identified and suspended all accounts likely belonging to the IRA. As we will show, new IRA accounts then popped up to take their place. Twitter identified and suspended these additional IRA accounts in June 2018. Twitter shared this information with the U.S. House of Representatives Permanent Select Committee on Intelligence (HPSCI), which released the names of these accounts to the public.¹⁵ Once the IRA accounts were suspended, all related information was no longer visible on Twitter— including the tweets themselves. That is, as of the writing of this document (January 2020), if one logs onto Twitter to check an IRA account, one simply sees a blank page.
- (Grey) Fortunately, Clemson University had been archiving Twitter. Clemson researchers went back into their archives and extracted the tweets from the suspended IRA accounts. They organized the tweets into 13 large spreadsheets and tagged each tweet with relevant metadata. On 31 July 2018, they published this dataset on the polling aggregation website *FiveThirtyEight*.^{16,17}

Meanwhile, back in the national security world:

- (Red) Machine learning researchers at IDA downloaded this dataset from *FiveThirtyEight* in early August 2018. This dataset consists of approximately 3 million tweets from approximately 3 thousand IRA accounts (specifically, 2,946,207 tweets from 2,848 IRA accounts). This is the dataset that was analyzed in this project. This dataset is a large subset of a similar dataset that Twitter later released to the public in October 2018¹⁸ and has continued to refine and supplement with tweets from similar information operations campaigns by organizations in other countries (e.g., Iran, Venezuela, Spain).¹⁹

-
- ⁵ Unified Combatant Commands. *Wikipedia Commons*.
https://commons.wikimedia.org/wiki/File:Unified_Combatant_Commands_map.png
- ⁶ Twitter Logo. *Wikipedia Commons*. https://commons.wikimedia.org/wiki/File:Twitter_logo.svg
- ⁷ *Report of the Select Committee on Intelligence, United States Senate, on Russian Active Measures Campaigns and Interference in the 2016 U.S. Election, Volume 2: Russia's Use of Social Media with Additional Views*. U.S. Senate Select Committee on Intelligence.
https://www.intelligence.senate.gov/sites/default/files/documents/Report_Volume2.pdf
- ⁸ Englund, Will. 2014 Mar 14. Kremlin Says Crimea Is Now Officially Part Of Russia After Treaty Signing, Putin Speech. *The Washington Post*. https://www.washingtonpost.com/world/russias-putin-prepares-to-annex-crimea/2014/03/18/933183b2-654e-45ce-920e-4d18c0ffec73_story.html
- ⁹ Kramer, Andrew E. and Gordon, Michael R. 2014 Aug 27. Ukraine Reports Russian Invasion On A New Front. *The New York Times*.
<https://www.nytimes.com/2014/08/28/world/europe/ukraine-russia-novoazovsk-crimea.html>
- ¹⁰ Hennigan, W.J.; Hennessey, Kathleen; and Mascaro, Lisa. 2014 Sep 22. U.S. Opens New Front Against Islamic State With Airstrikes In Syria. *Los Angeles Times*. <https://www.latimes.com/world/middleeast/la-fg-us-syria-isis-airstrikes-20140922-story.html>
- ¹¹ Hall, Allan and Lichfield, John. 2015 Aug 24. Germany Opens Its Gates: Berlin Says All Syrian Asylum-Seekers Are Welcome To Remain, as Britain Urges To Make A "Similar Statement". *Independent*. <https://www.independent.co.uk/news/world/europe/germany-opens-its-gates-berlin-says-all-syrian-asylum-seekers-are-welcome-to-remain-as-britain-is-10470062.html>
- ¹² Gordon, Michael R.; Cooper, Helene; and Shear, Michael D. 2017 Apr 06. Dozens Of U.S. Missiles Hit Air Base In Syria. *The New York Times*. <https://www.nytimes.com/2017/04/06/world/middleeast/us-said-to-weigh-military-responses-to-syrian-chemical-attack.html>
- ¹³ Barret, Devlin; Horwitz, Sari; and Helderman, Rosalind S. 2018 Feb 16. Russian Troll Farm, 13 Suspects Indicted in 2016 Election Interference. *The Washington Post*. https://www.washingtonpost.com/world/national-security/russian-troll-farm-13-suspects-indicted-for-interference-in-us-election/2018/02/16/2504de5e-1342-11e8-9570-29c9830535e5_story.html
- ¹⁴ *United States of America v. Internet Research Agency LLC*. U.S. District Court for the District of Columbia, No. 18 U.S.C. 2 371 1349 1028A. 2018 Feb 16. <https://www.justice.gov/file/1035477/download>
- ¹⁵ *Exposing Russia's Effort to Sow Discord Online: The Internet Research Agency and Advertisements*. U.S. House of Representatives Permanent Select Committee on Intelligence. <https://intelligence.house.gov/social-media-content/default.aspx>
- ¹⁶ Roeder, Oliver. 2018 Jun 2013. Why We're Sharing 3 Million Russian Troll Tweets. *FiveThirtyEight*.
<https://fivethirtyeight.com/features/why-were-sharing-3-million-Russian-troll-tweets/>
- ¹⁷ Linvill, Darren L. and Warren, Patrick L. 2018. *Troll Factories: The Internet Research Agency and State-Sponsored Agenda Building*.
http://pwarren.people.clemson.edu/Linvill_Warren_TrollFactory.pdf
- ¹⁸ Gadde, Viljaya and Roth, Yoel. 2018 Oct 17. *Enabling Further Research Of Information Operations On Twitter*.
https://blog.twitter.com/en_us/topics/company/2018/enabling-further-research-of-information-operations-on-twitter.html
- ¹⁹ Roth, Yoel. 2019 Jun 13. *Information Operations On Twitter: Principles, Process, and Disclosure*.
https://blog.twitter.com/en_us/topics/company/2019/information-ops-on-twitter.html

IDA | IDA did not...

- IDA did **not** create a “Russian troll detector”
 - Would have also needed a control dataset of tweets that were truly **not** posted by Russian trolls → Type I **and** II errors, FNs **and** FPs, Pd **and** Pfa, Recall **and** Precision
 - May not have been useful anyway – trolls are not unique to Russians
- IDA did **not** jump straight into supervised learning
 - So far, IDA has used unsupervised learning to **understand** the IRA tweets
 - What kind of detector **could** be created from this dataset?
 - What kind of detector **could** be useful to intelligence analysts?
 - Next, IDA must use this understanding to create an **analyst dashboard** that can **adapt itself over time**, for new tweets

Before we discuss what we *did* do in this analysis, we should first make clear what we *did not* do.

We did *not* create a “Russian troll detector”. To do that, we would have needed both:

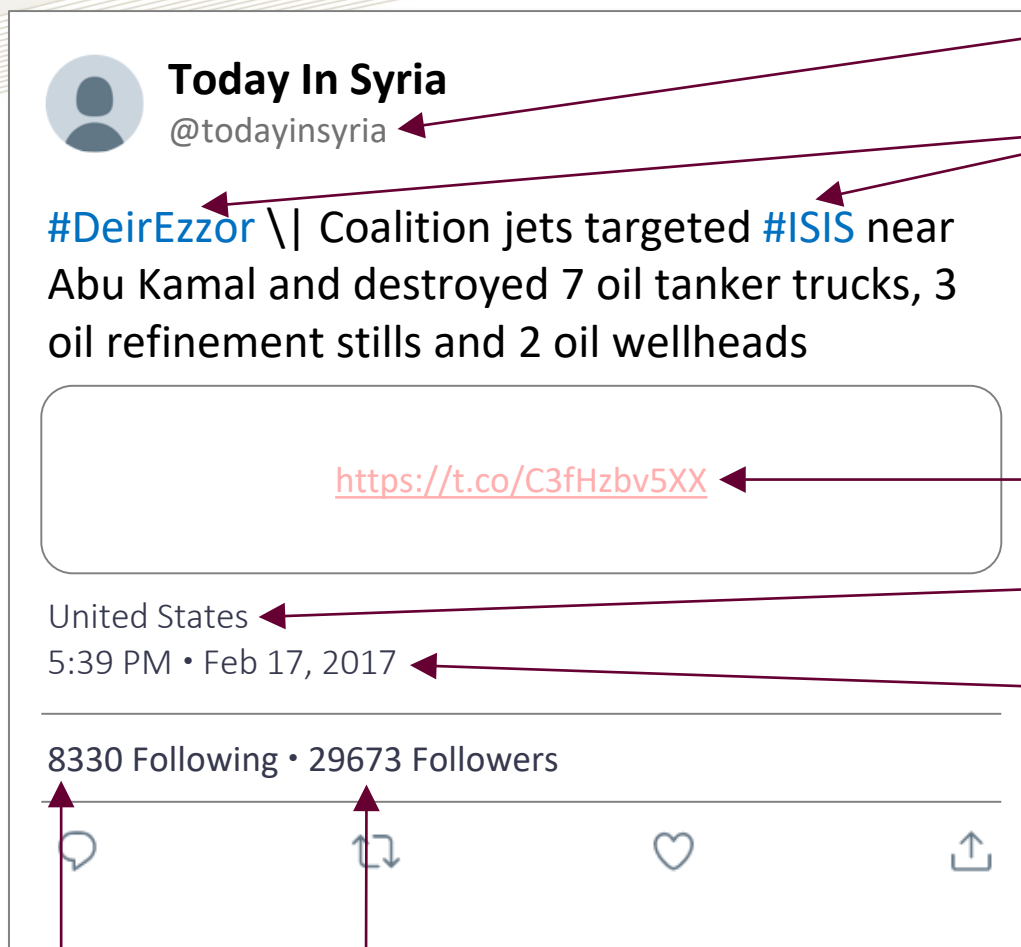
- A dataset of IRA tweets (i.e., tweets truly posted by Russian trolls), such as the Clemson dataset described on the previous slide *and*
- A control dataset of tweets posted by regular Twitter users (i.e., tweets truly posted by *non*-Russian trolls)

Only with both datasets could one score how well a system can accurately detect Russian trolls *and* accurately ignore regular Twitter users, using performance metrics to quantify Type I *and* Type II errors, False Negatives (FNs) *and* False Positives (FPs), Recall *and* Precision, and so forth.

Furthermore, we did *not* jump straight into supervised machine learning. Instead, we have so far used unsupervised learning to *understand* the IRA tweets, exploring questions such as:

- What kind of detector *could* be created from this dataset?
- What kind of detector *could* be useful to intelligence analysts?

We must use this understanding in future work to create an analyst dashboard that can adapt itself over time, for new tweets, as we discuss at the end of this document.



Handle

Hashtags

Content

URL to other linked content

Region

Time and Date

8330 Following • 29673 Followers

Following

Followers

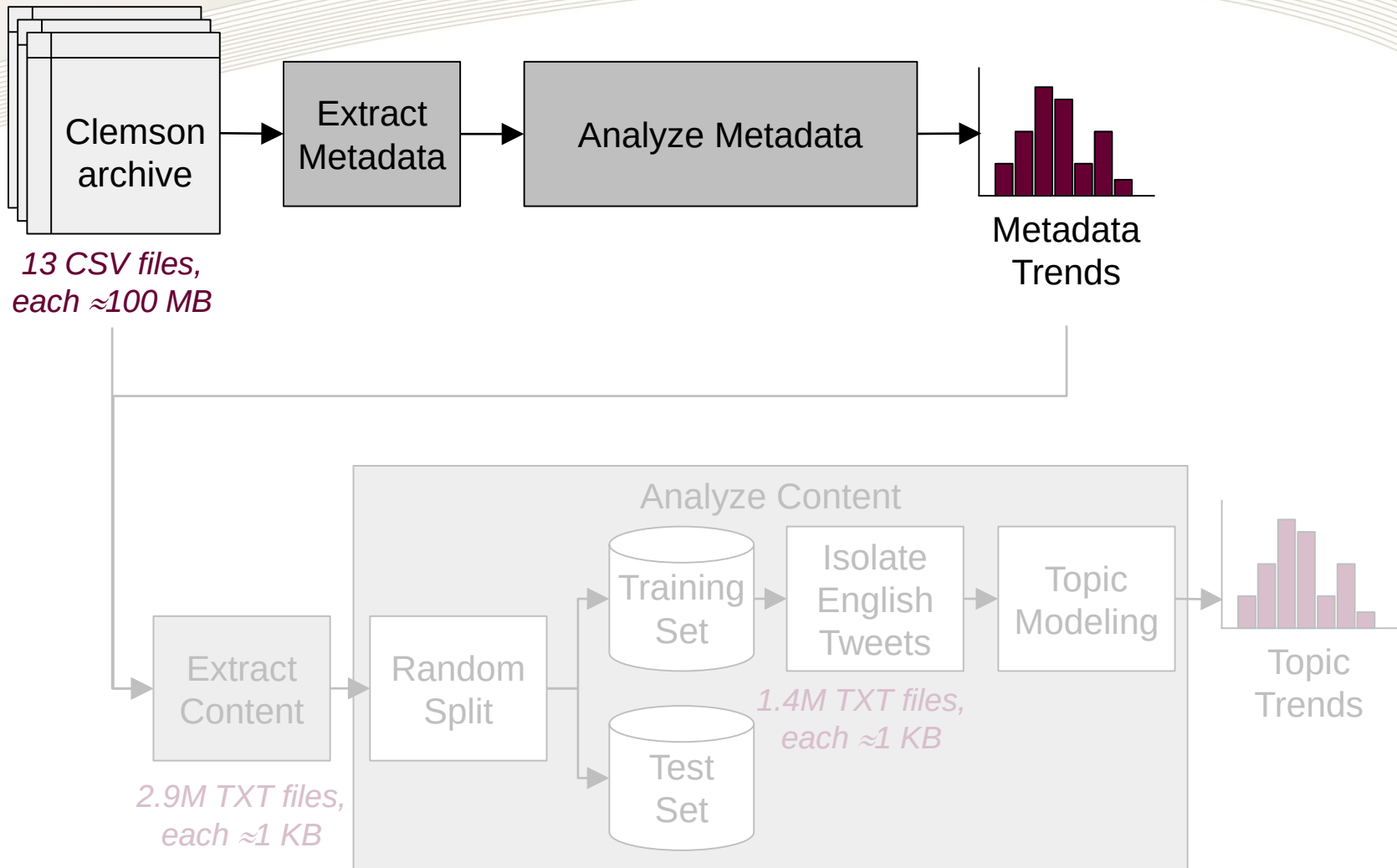
We now describe the anatomy of a tweet.

Because Twitter has suspended the IRA accounts, they are no longer available to view online. Therefore, we show a mock-up example of how one IRA tweet would have displayed to a Twitter user. The information shown in this example has been taken from a suspended IRA tweet.

- First there is the handle, which denotes the name of the Twitter account. All handles begin with the @ symbol. This example tweet was posted by the handle @todayinsyria.
- Then the content of the tweet is displayed, which can contain different types of data: text, hashtags, URLs, images, videos, and so forth. Hashtags begin with the # symbol and are a poster's self-labeling mechanism to describe or tag the tweet according to a trending topic. In this example, the @todayinsyria poster chose to self-label this tweet with the hashtags #DeirEzzor and #ISIS. Furthermore the poster can also link websites, images, and video to a tweet. For example, the @todayinsyria poster linked this tweet to a website using the website's URL: <https://t.co/C3fHzbv5XX>. In most web browsers, Twitter will automatically fetch an image or video from the linked website and render that in place of the URL itself, to display along with the tweet.
- Each tweet is time- and date-stamped. This tweet was posted at 5:39 PM on 17 February 2017.
- If the handle has its location settings turned on, then the poster's supposed geographical location at the time he or she posted the tweet is also reported. In this example, @todayinsyria's location settings were turned on and the @todayinsyria poster was supposedly physically located in the United States at the time he or she posted the tweet. (More likely, though, he or she was spoofing his or her account to only make it *appear* as though he or she were physically located in the United States, as we will discuss later.)

- Finally, the handle's followings and followers are displayed:
 - Followings are the accounts that the handle is following– when the @todayinsyria poster logged onto Twitter, his or her Twitter feed automatically displayed the tweets recently posted by all 8,330 of his or her followings.
 - In contrast, followers are the accounts that follow the handle– there were 29,673 accounts whose posters saw @todayinsyria's tweets in their Twitter feeds.

This page is intentionally left blank.



IDA first focused on the tweet metadata.

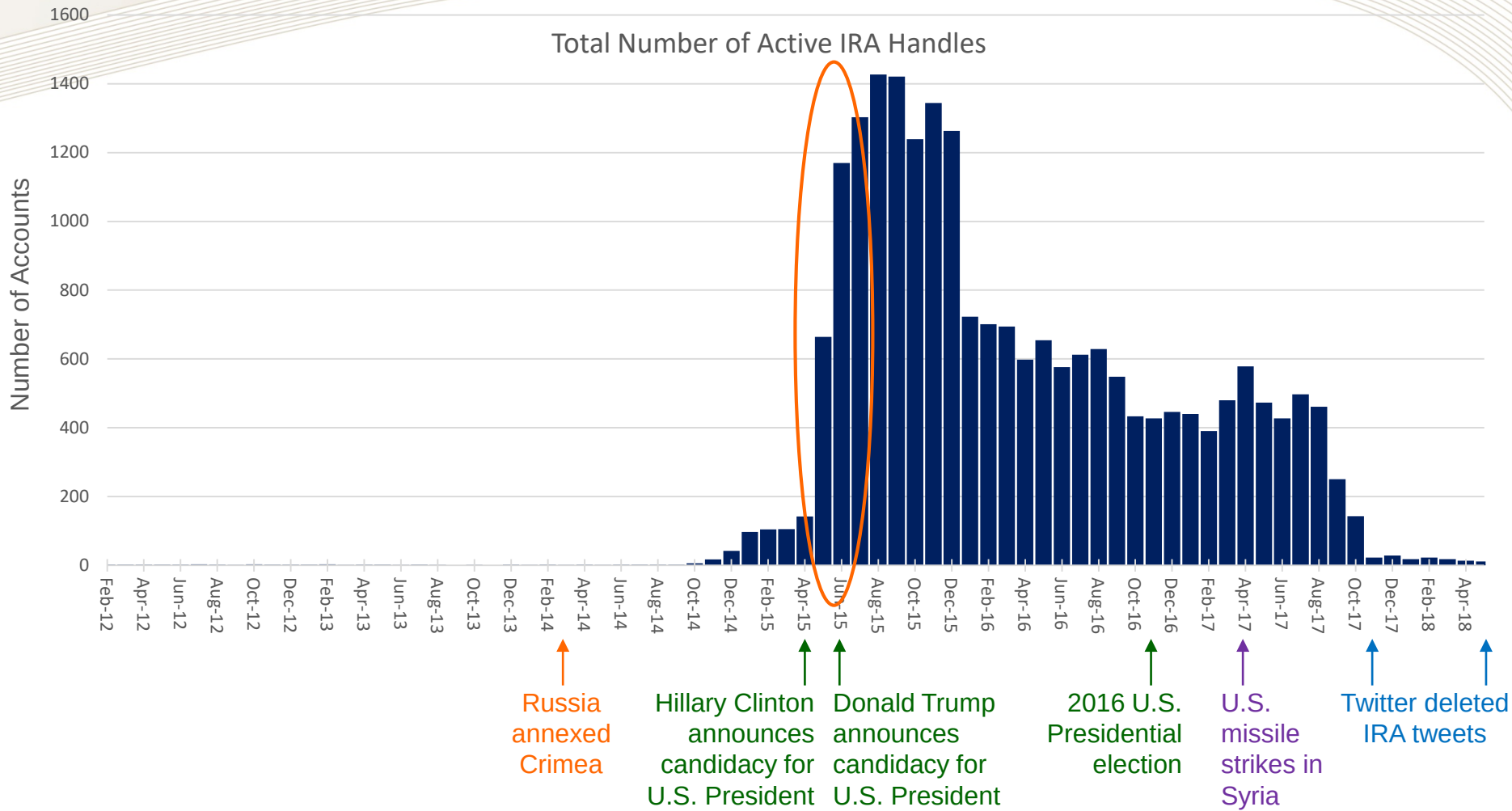
The Clemson dataset consisted of a large number of small pieces of data– approximately 3 million tweets, each fewer than 280 characters in length. These tweets were organized into 13 large comma-separated value (CSV) files, each approximately 100 MB. We developed a processing pipeline to efficiently analyze these tweets. The processing pipeline consisted of two main steps, focused on the tweets' (1) metadata and (2) content.

In the first step of our analysis, we extracted and analyzed the tweets' *metadata*. As discussed earlier, the Clemson researchers had tagged each tweet with associated metadata such as handle, time and date, location, and so forth. This metadata was a rich source of information, all by itself.

IDA | IDA explored the IRA tweet metadata

- **Handle:** The name of the IRA account
- **Followers:** Number of accounts that are following the IRA handle
- **Following:** Number of accounts the IRA handle follows
- **Region:** The country the IRA handle (supposedly) posted the tweet from
- **Language:** The language the IRA handle posted the tweet in
- **Category:** Clemson's manually-applied label for the IRA handle, based on all of its tweets' contents (including hashtags, URLs, etc)

We first considered the IRA handles (i.e., the IRA accounts, rather than the individual tweets).



IRA started in early 2012 (or even before), but ramped up in May/June 2015

We plot the number of active IRA handles (accounts) versus time. The number of active handles are binned by month, with some events of interest noted beneath the plot. An “active” handle is one that posted at least one tweet in a given month. The very first IRA tweet in this dataset was posted in early 2012. The IRA began tweeting from notably more handles in late 2014. There was a significant increase in the number of active handles between April and June 2016 (an 823% increase). There was also a notable decline in active handles in November 2017, when Twitter first suspended the IRA accounts. The number of active handles did not drop completely to zero, however, showing that the IRA was able to create new accounts fairly quickly. The dataset ends in May 2018, upon the second Twitter suspension.

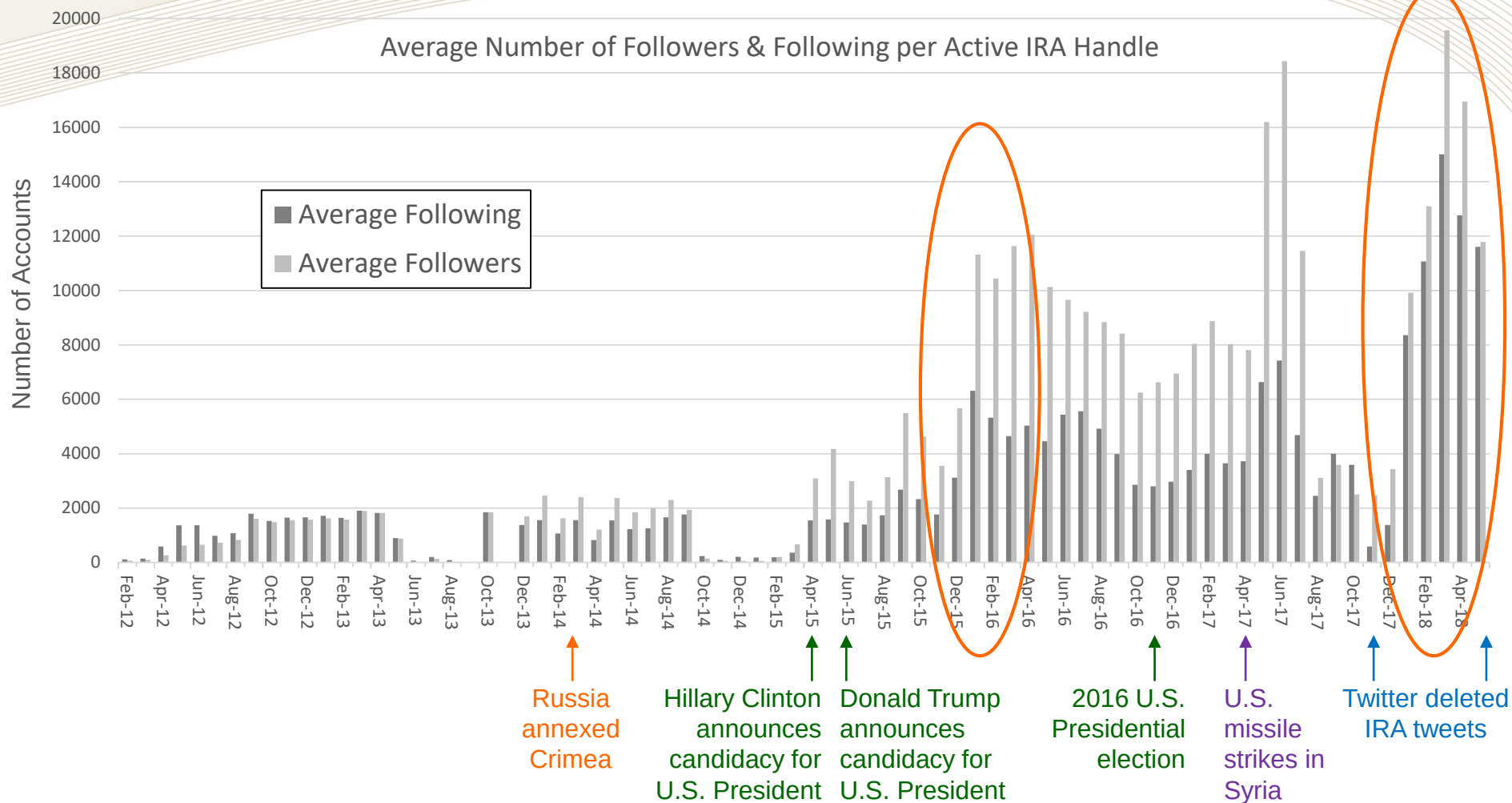
- **Handle:** The name of the IRA account
- **Followers:** Number of accounts that are following the IRA handle
- **Following:** Number of accounts the IRA handle follows
- **Region:** The country the IRA handle (supposedly) posted the tweet from
- **Language:** The language the IRA handle posted the tweet in
- **Category:** Clemson's manually-applied label for the IRA handle, based on all of its tweets' contents (including hashtags, URLs, etc)

IDA then considered the IRA's followings and followers. As discussed on a previous slide:

- Followings are the accounts that the IRA handle is following– when the IRA poster logs on to Twitter, he or she sees the tweets recently posted by his or her followings.
- In contrast, followers are the accounts that follow the IRA handle– the posters of these accounts see the IRA tweets when they log on to Twitter.



IDA investigated IRA followers & followings over time



After first Twitter deletion in Nov 2017,
each new IRA account quickly established an audience

We plot the average number of following and follower accounts per active IRA handle versus time. On average, IRA accounts had more followers (light grey) than followings (dark grey). That is, more accounts (light grey) followed the IRA than the IRA followed other accounts (dark grey).

The smaller orange circle points out the large increase in followings (dark grey) and especially followers (light grey) that took place, on average, for each active IRA account in early 2016– which we will discuss later.

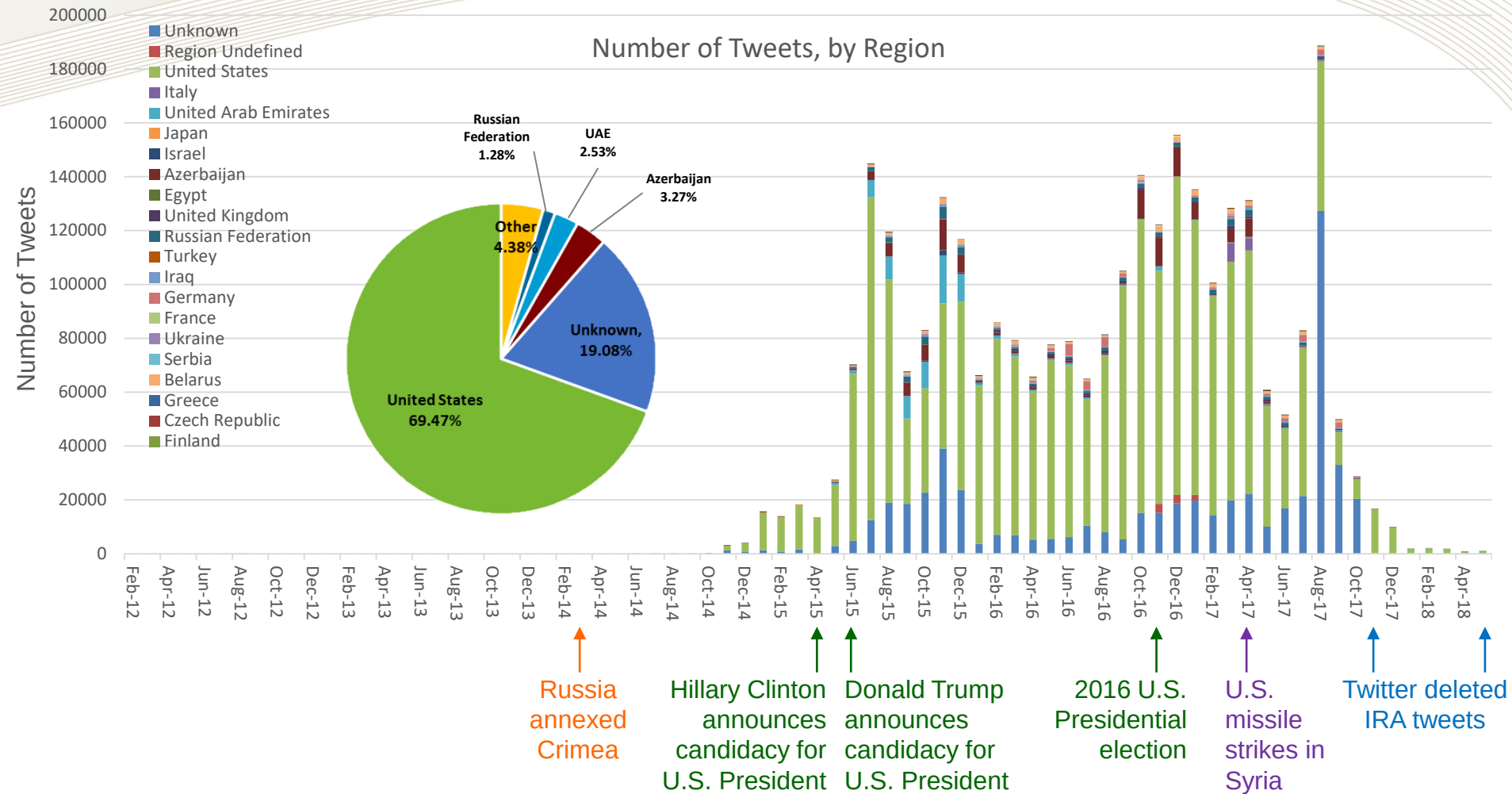
The larger orange circle points out that, after the first Twitter suspension in November 2017, each new IRA account quickly established, on average, a large number of both followings and followers.

IDA | IDA explored the IRA tweet metadata

- **Handle:** The name of the IRA account
- **Followers:** Number of accounts that are following the IRA handle
- **Following:** Number of accounts the IRA handle follows
- **Region:** The country the IRA handle (supposedly) posted the tweet from
- **Language:** The language the IRA handle posted the tweet in
- **Category:** Clemson's manually-applied label for the IRA handle, based on all of its tweets' contents (including hashtags, URLs, etc)

IDA also explored the locations (i.e., geographical regions), from which the tweets were supposedly posted. Although it is now known that most, if not all, of these tweets were actually posted from the IRA headquarters in St. Petersburg, Russia Federation, the IRA posters were spoofing their IP addresses to make it *appear* as though they were tweeting from all over this world.²⁰ Thus the information on the next slide reflects the geographical regions that the IRA wanted their followers and followings to *think* they were posting from. Presumably, these were the geographical regions that the IRA was targeting.

²⁰ *Report of the Select Committee on Intelligence, United States Senate, on Russian Active Measures Campaigns and Interference in the 2016 U.S. Election, Volume 2: Russia's Use of Social Media with Additional Views*. U.S. Senate Select Committee on Intelligence.
https://www.intelligence.senate.gov/sites/default/files/documents/Report_Volume2.pdf



Most IRA tweets were (supposedly) posted from the U.S.
Other prominent regions were Azerbaijan, UAE, Russian Federation, etc.

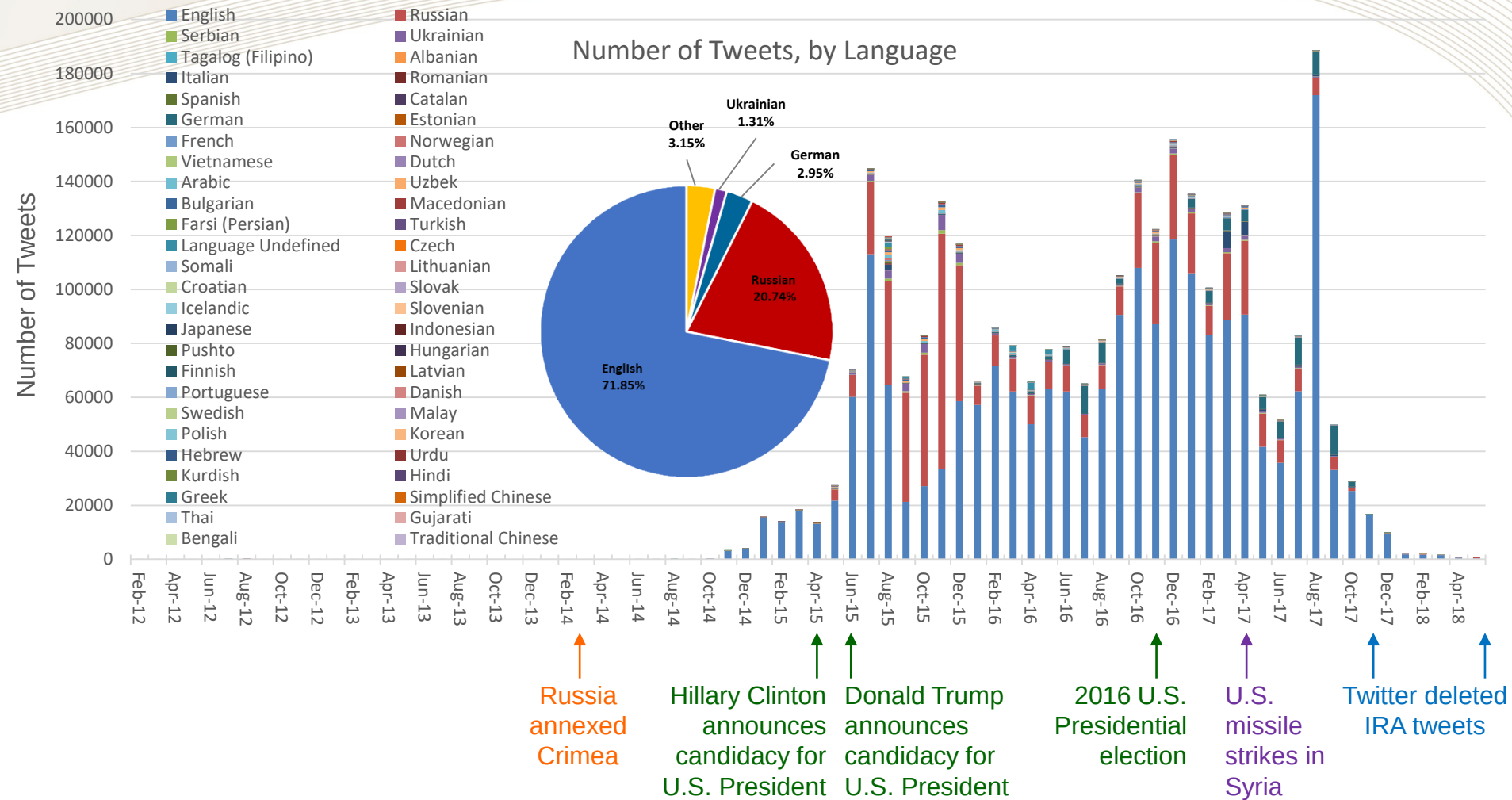
We plot the number of IRA tweets versus time. As before, the horizontal axis shows time, binned by month. However, unlike previous slides, the vertical axis now shows the number of IRA *tweets*, rather than the number of accounts. The tweets are separated by color according to the geographical region from which the IRA wanted them to appear to be posted from. Most IRA tweets were supposedly posted from the United States (69.47%, green), presumably indicating that the IRA targeted Twitter users residing in the United States. Other prominent regions were Azerbaijan (3.27%, red), the United Arab Emirates (UAE) (2.53%, light blue), and the Russian Federation (1.28%, dark blue). All other regions collectively accounted for 4.38% (yellow) of the total number of tweets. A large subset of the IRA tweets (19.08%, medium blue) were tweeted from an unknown region— perhaps because the IRA handles did not have their location services turned on.

IDA | IDA explored the IRA tweet metadata

- **Handle:** The name of the IRA account
- **Followers:** Number of accounts that are following the IRA handle
- **Following:** Number of accounts the IRA handle follows
- **Region:** The country the IRA handle (supposedly) posted the tweet from
- **Language:** The language the IRA handle posted the tweet in
- **Category:** Clemson's manually-applied label for the IRA handle, based on all of its tweets' contents (including hashtags, URLs, etc)

We also explored the languages in which the tweets were posted.

IDA explored the languages the IRA tweets were posted in



Most IRA tweets were posted in English.
Other prominent languages were Russian, German, Ukrainian, etc. 38

Once again, we plot the number of IRA tweets versus time. The shape of this histogram is identical to that shown in the previous plot for geographical region. This time, though, the color denotes the language that the tweet was posted in. Most of the tweets were posted in English (71.85%, medium blue), presumably indicating that the IRA targeted an English-speaking audience. Other prominent languages were Russian (20.74%, red), German (2.95%, dark blue), and Ukrainian (1.31%, purple). All other languages collectively comprised 3.15% (yellow) of the total.

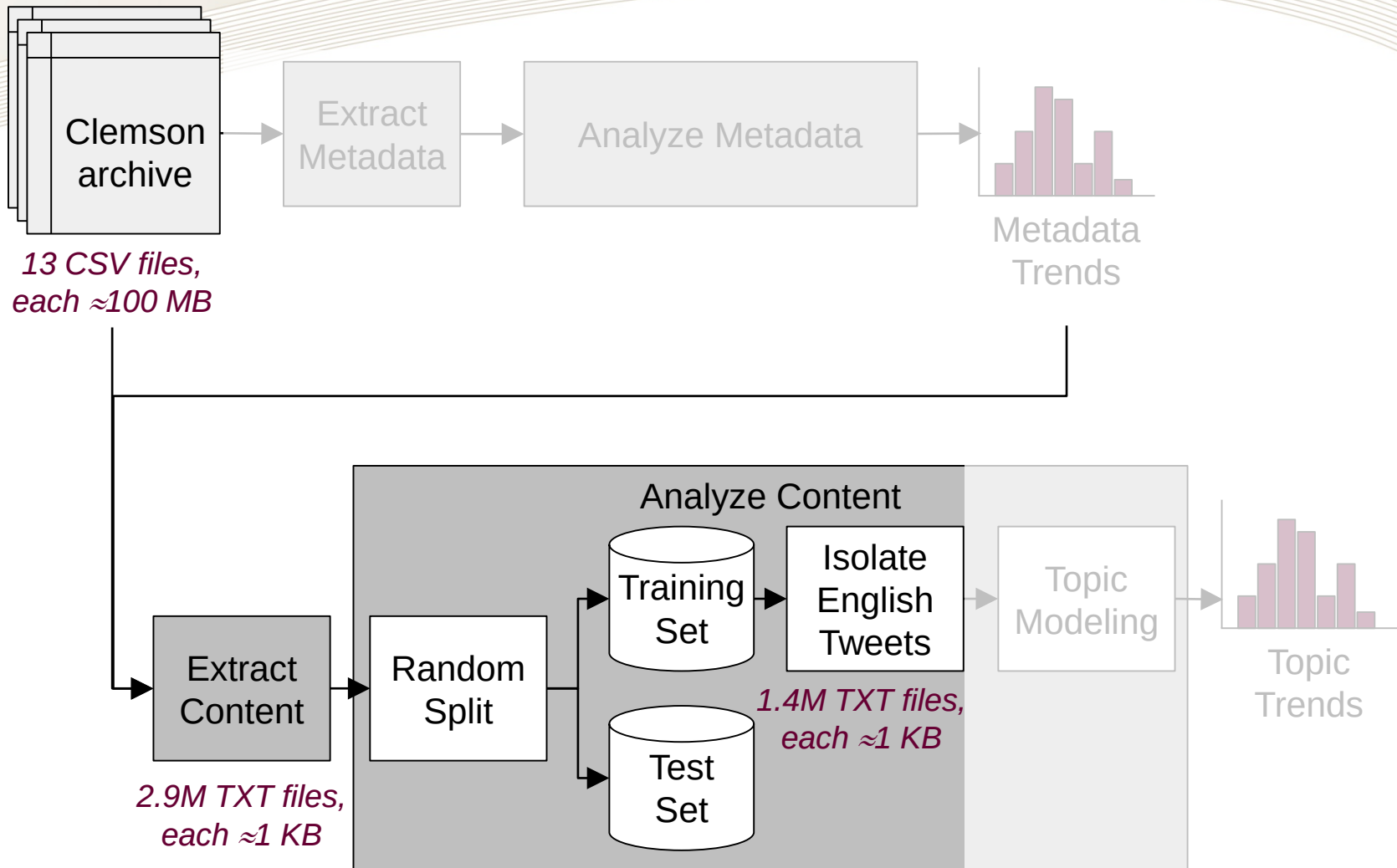
IDA | IDA explored the IRA tweet metadata

- **Handle:** The name of the IRA account
- **Followers:** Number of accounts that are following the IRA handle
- **Following:** Number of accounts the IRA handle follows
- **Region:** The country the IRA handle (supposedly) posted the tweet from
- **Language:** The language the IRA handle posted the tweet in
- **Category:** Clemson's manually-applied label for the IRA handle, based on all of its tweets' contents (including hashtags, URLs, etc)

Existing metadata have limitations when inferring intent:
e.g., just because a tweet was supposedly posted in the U.S. doesn't mean it was posted by an American.

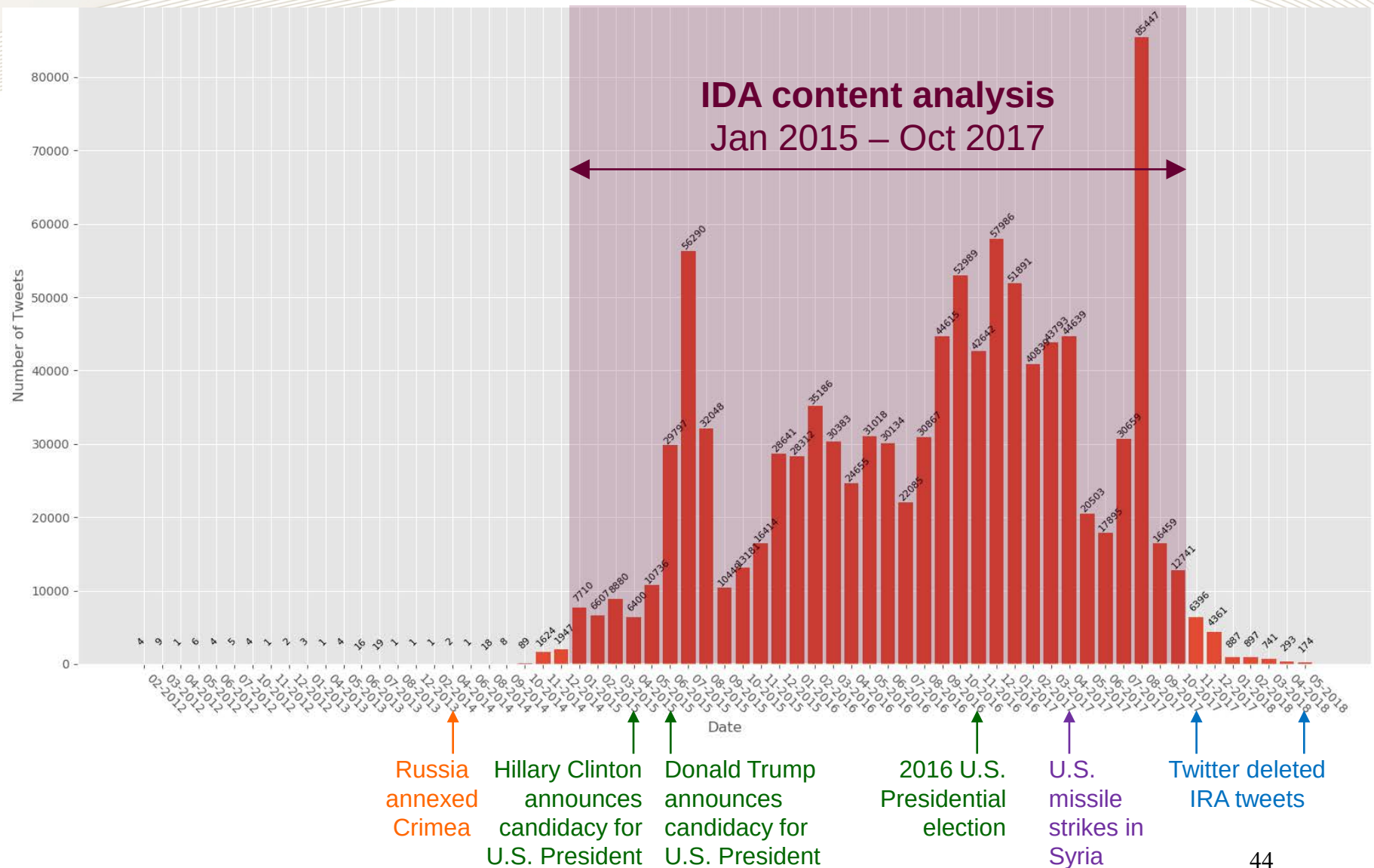
We must now analyze the content of each tweet – in an automated manner.

Trends in metadata are useful in identifying the audiences targeted by the IRA— namely, an English-speaking audience in the United States. However, the metadata has inherent limitations in inferring the IRA’s intent— the messages that the IRA was trying to convey to its audiences. To better understand the IRA’s intent, one must analyze the *content* of their tweets. To aid in this effort, the Clemson researchers provided one final piece of metadata— a six-category scheme to summarize the overall theme of each IRA handle (i.e., left troll, right troll, fearmonger, commercial, hashtag gamer, and news feed). Trends in these categories are described in the backup slides. Unfortunately, these categories varied only by handle, not by individual tweet. Manually reading and tagging each individual tweet would be a daunting task. Therefore, automated or at least semi-automated methods were needed to explore the content of the individual tweets.



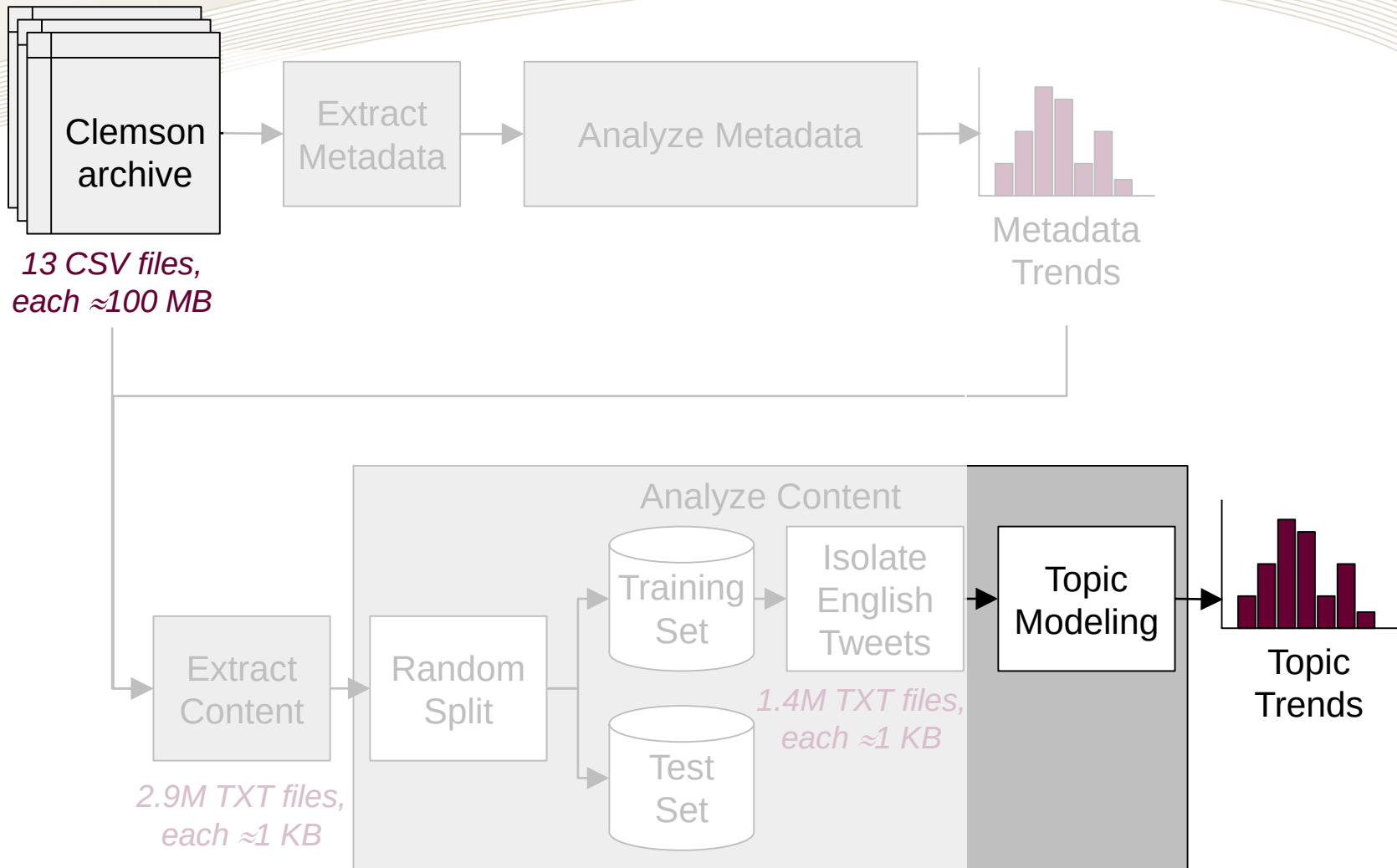
IDA later focused on the tweet content.

In the next step of our analysis, we extracted and analyzed the tweets' content. As discussed in previous slides, the Clemson researchers had organized the tweets into 13 CSV files, each approximately 100 MB. Our pipeline converted those CSV files into approximately 2.9 million text files, each containing the content of a single, individual tweet, approximately only 1 KB in size. We then randomly split the tweets 50/50 into a training and test set, in anticipation of future work in which we may use supervised learning techniques that will require separate training and test sets. We focused the remaining steps of our analysis on the training set only. We also focused on only the English tweets, since our metadata analysis showed that the IRA posted most of their tweets in English.



We plot the number of English tweets in the training set over time. As in previous slides, the horizontal axis shows time while the vertical axis shows number of tweets. The number of tweets in any given month are written in small black text above each bar of the histogram. For example, there were 7,710 training set tweets posted in English in January 2015 and 85,447 posted in August 2017.

We sought to understand how the *patterns* in the IRA's tweet topics *changed over time*. To do this, we treated each month of English training set tweets as its own individual corpus, exploring which topics the IRA tweeted about in each individual month. To do this in as automated a fashion as possible, one must fit a statistical model to each corpus of tweets, and the number of tweets in that corpus must be large enough to allow the parameters of the statistical model to converge. Therefore we focused our content analysis on only those months for which a large number of training set tweets were posted—namely, January 2015 through October 2017 (the month preceding the first Twitter suspension).



IDA later focused on the tweet content.

The final step of our pipeline was to perform our content analysis– specifically, to perform *topic modeling* on each month’s corpus of tweets. Topic modeling is a semi-automated process in which a statistical model is fit to a corpus of documents (i.e., a month of tweets) in order to determine the main topics contained in the documents. Topic modeling is a form of unsupervised machine learning.

- LDA fits a statistical model to a corpus of documents, clustering the main topics in each:
 - **Corpus** = month's worth of tweets, e.g. Feb 2017
 - **Document** = tweet content, e.g. #DeirEzzor \| Coalition jets targeted #ISIS near Abu Kamal and destroyed 7 oil tanker trucks, 3 oil refinement stills and 2 oil wellheads
 - **Topic** = list of associated words, e.g. isis syria targeted targets Iraq forces accounts Israel u.s mosul opiceisis iceisis aleppo north killed yemen refugees airport Syrian saa
- A human must then label the topics:
 - **Topic Label** = 1-2 word summary of topic, e.g. Syrian Conflict



We used a well-known approach called Latent Dirichlet Allocation (LDA) to perform topic modeling. LDA fits a statistical model to a corpus of documents, clustering the main topics.²¹ In our analysis:

- A *corpus* consists of one month's worth of tweets (e.g., February 2017). The statistical assumption behind LDA is that each corpus can be summarized by a fixed number of topics.
- A *document* is the content of a single tweet (e.g., #DeirEzzor \ | Coalition jets targeted #ISIS near Abu Kamal and destroyed 7 oil tanker trucks, 3 oil refinement stills and 2 oil wellheads). LDA assumes that each document in the corpus is composed of words drawn from a mixture of the topics.
- A *topic* is a list of associated words (e.g., isis syria targeted targets Iraq forces accounts Israel u.s mosul opiceisis iceisis aleppo north killed yemen refugees airport Syrian saa).

LDA is implemented in a variety of software packages and code libraries. We used MALLET to perform LDA. MALLET is a well-known, free software package written in Java²² and can be quickly and easily installed and used on any Windows or Mac computer with administrator privileges.

For the purposes of our analysis, and the ease of our readers, we chose to summarize each topic with a topic label. In our analysis:

- A *topic label* is a 1-4 word summary of a topic (e.g., Syrian Conflict).

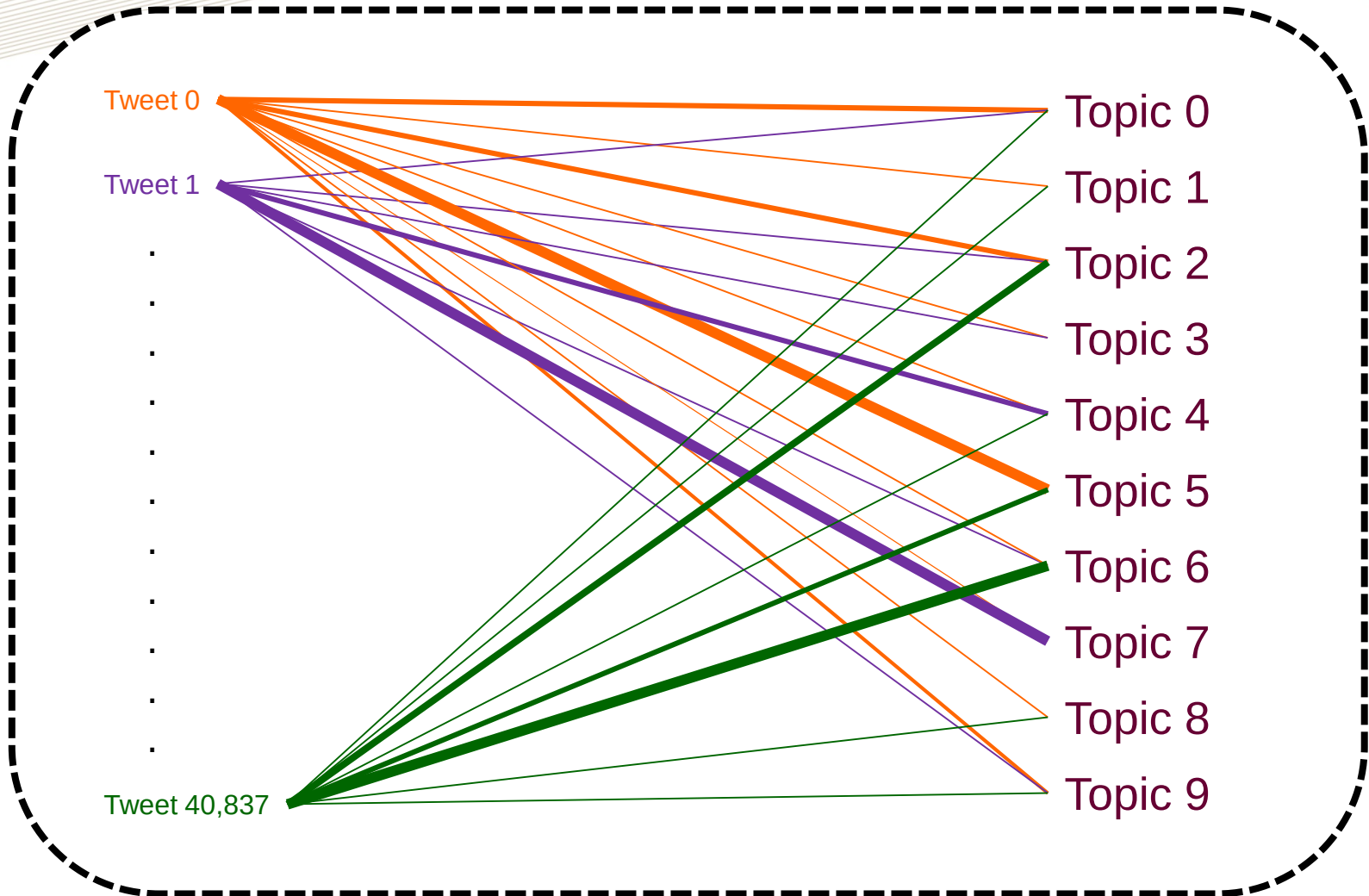
Human intelligence is needed to manually perform this final, topic labeling step. To accurately interpret and summarize the long lists of words for each topic, the human labeler requires expertise in (1) the language in which the tweets were written (in this case, English) and (2) the social and cultural themes that Twitter users were likely to have encountered during the period of analysis (in this case, in the United States from January 2015 to October 2017). We served as our own human topic labelers for the purposes of this analysis.

²¹ Blei, David M.; Ng, Andrew Y.; and Jordan, Michael I. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning*, 3: 993-1022.
<http://jmlr.org/papers/volume3/blei03a/blei03a.pdf>

²² MALLET Machine Learning for Language Toolkit. <http://mallet.cs.umass.edu/>

This page is intentionally left blank.

Month

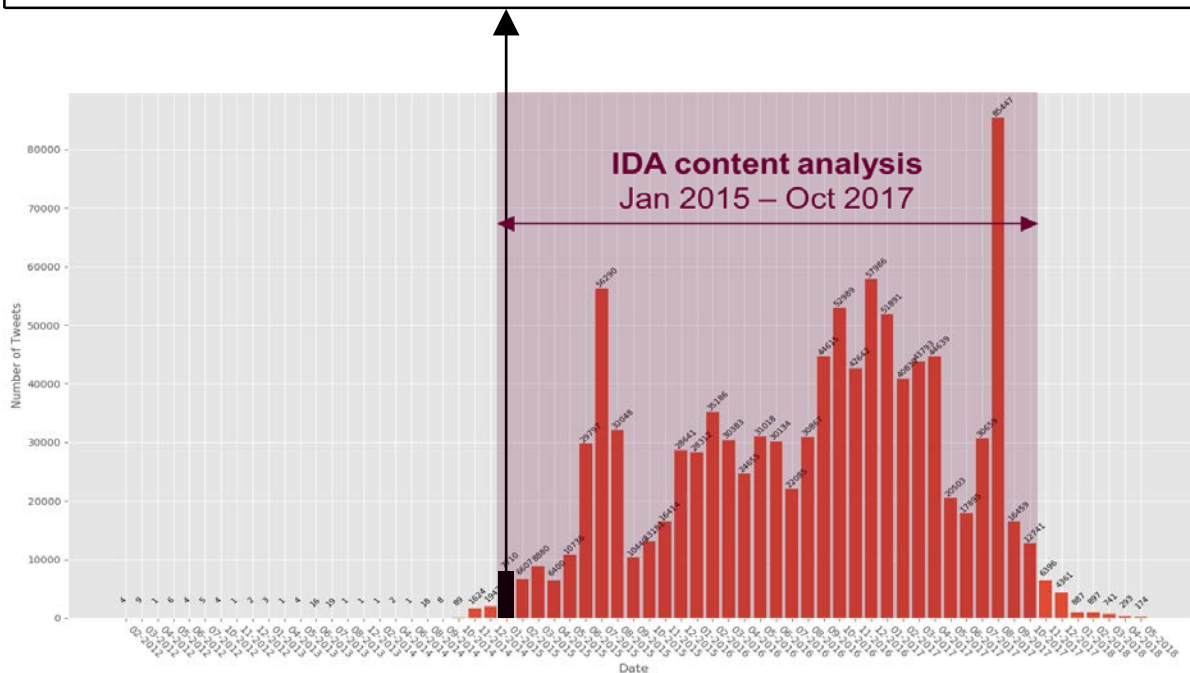


LDA allows each document (tweet) to have multiple topics.

This slide shows a visual representation of LDA. In this example, there are 40,838 tweets (documents) in a month (corpus), ranging from Tweet 0 through Tweet 40,837. The LDA statistical model assumes that there were 10 topics in the month, ranging from Topic 0 through Topic 9. Any individual tweet does not necessarily “belong” to a single topic and could be associated with multiple topics. LDA quantifies the strength of a tweet’s association to each topic, represented here as the thickness of each tweet-to-topic line. For example, the final Tweet 40,837 is most strongly associated with Topic 6, followed by Topic 2. This final tweet is hardly associated with Topics 3 or 7 at all.

Topics (Automated)

0	0.03219	life love heart you're god iamonfire don't people make can't true li...
1	0.01375	quotes love quote iamonfire man movie book live entertainment time l...
2	0.02503	local news miami man police fire found topnews dead shooting crash h...
3	0.04037	don't love life people make time iamonfire things good smile day hea...
4	0.01442	oscarformuhammad oscarnoms politics future whoismuhammad power past ...
5	0.01855	love iamonfire life learn man great william art money he's james yea...
6	0.02446	news politics worldnews cbs obama state topvideo u.s topnews arizona...
7	0.01233	entertainment fukushimaagain news chernobyl ukraine nuclear someonev...
8	0.0209	sports news local seahawks bowl super game state team nfl snow patri...
9	0.01302	make things trust fear didn't shit sense money confidence lol build ...



This slide shows the output of the topic model that we fit to the English training set tweets posted in the first month of our analysis, January 2015 (colored in black in the histogram below).

- The first column refers to the topic number. To automate our processing pipeline as much as possible, we fit our LDA statistical model to each month using the same number of topics (10) for each month. We hypothesized that the IRA management likely instructed their posters to tweet about no more than 10 different topics on any given month. As we will show, there is evidence that 10 was not always the correct number of topics for each month. However, 10 topics still allowed us get a sense of the overall pattern of topics that the IRA was tweeting about in a given month. Thus the specific number of topics in a given month (10 vs. 9 vs. 8) was less important than the overall pattern of topics in that month.
- The second column refers to the strength of the topic in the month, with larger numbers indicating a stronger topic.
- The final column contains the words most strongly associated to that topic.

For instance, when fitting a 10-topic model, the strongest topic in January 2015 was Topic 3, for which the most strongly associated words were: don't, love, life, people, make, time, iamonfire, things, good smile, day, and so forth.

Topics (Automated)

```

0 0.03219 life love heart you're god iamonfire doi...
1 0.01375 quotes love quote iamonfire man movie b...
2 0.02503 local news miami man police fire found 1...
3 0.04037 don't love life people make time iamonf:...
4 0.01442 oscarformuhammad oscarnoms politics futi...
5 0.01855 love iamonfire life learn man great wil...
6 0.02446 news politics worldnews cbs obama state...
7 0.01233 entertainment fukushimaagain news chern...
8 0.0209 sports news local seahawks bowl super ga...
9 0.01302 make things trust fear didn't shit sense...

```

Topic Labels (Human)

```

0 Feel Good/Motivation
1 Books/Film
2 Local Headlines/Crime
3 Feel Good/Motivation
4 Iranian Film
5 Feel Good/Motivation
6 International Headlines
7 Nuclear Issues
8 Sports
9 Feel Good/Motivation

```

**Visual
assessment:**
Highlight
words in each
topic with
similar
semantic
meaning

January 2015

0 life love heart you're god iamonfire don't people make can't true live dreams feel leave world ihq dream find lose

1 quotes love quote iamonfire man movie book live entertainment time life action dark light king bookquotes secret prayer moviequotes david

2 local news miami man police fire found topnews dead shooting crash home woman car county crime shot arrested suspect accused

3 don't love life people make time iamonfire things good smile day heart success give bad you're thing hate hard makes

4 oscarformuhammad oscarnoms politics future whoismuhammad power past peace oscarssowhite mind people news prophet oil men health i'm happiness local muhammad

5 love iamonfire life learn man great william art money he's james years die eat answer bitch big ain't beautiful live

6 news politics worldnews cbs obama state topvideo u.s topnews arizona republic local court union terrorism house bill press paris storm

7 entertainment fukushimaagain news chernobyl ukraine nuclear someonewhokillschildren president seattle cat time american government npp ukrainian years that's disaster big care

8 sports news local seahawks bowl super game state team nfl snow patriots fan top football wins win washington fans season

9 make things trust fear didn't shit sense money confidence lol build dog can't human small real happy cats body cute

**Manual
label:**
Assign label
to each topic,
based on
highlighted
words

We then manually summarized each topic with a 1-4 word topic label. To do this, we printed out all topic word lists from all months of our analysis period and physically spread them out over a large surface so that we could manually read all of them in short succession:

- We visually assessed each topic word list, highlighting words in each topic with a similar semantic meaning. Highlighted colors indicated a potential topic label, to help us qualitatively and subjectively determine if the same topic was represented in multiple month's corpuses.
- Then, we manually labeled each topic word list, assigning a label to each topic based on the highlighted words. For example, Topic 0 (life love heart you're god iamonfire...) was labeled "Feel Good/Motivation", while Topic 1 (quotes love quote iamonfire man movie book...) was labeled "Books/Film" and Topic 2 (local news miami man police fire found) was labeled "Local Headlines/Crime".

Topics (Automated)

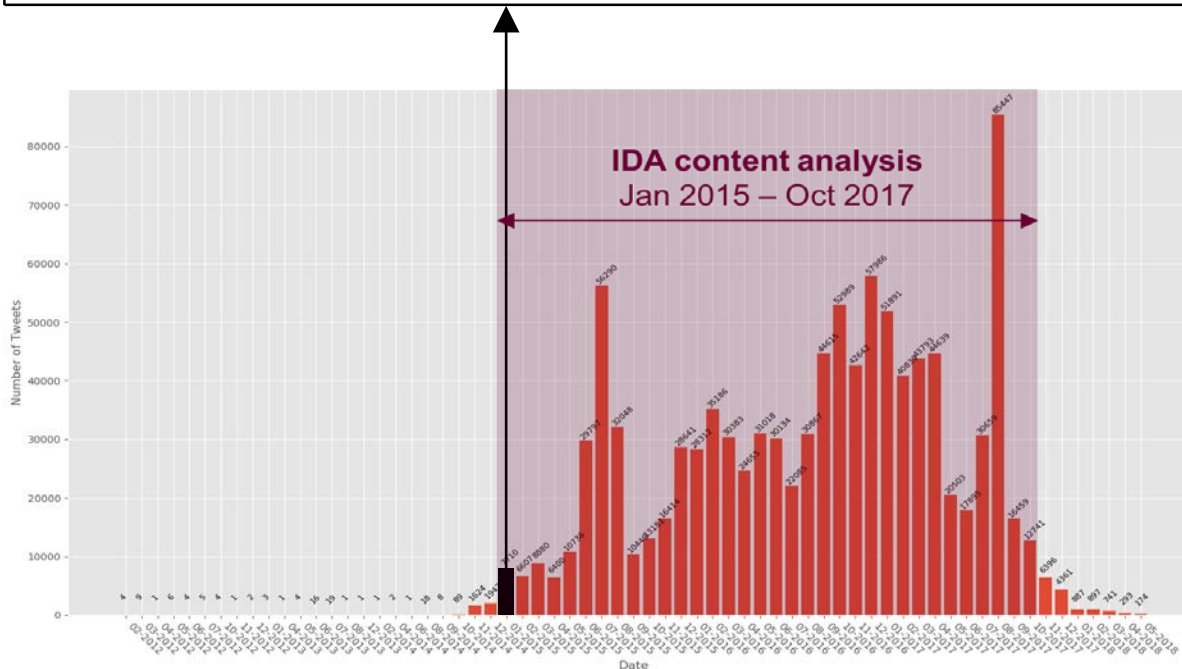
```

0 0.03219 life love heart you're god iamonfire doi...
1 0.01375 quotes love quote iamonfire man movie b...
2 0.02503 local news miami man police fire found 1...
3 0.04037 don't love life people make time iamonf:...
4 0.01442 oscarformuhammad oscarnoms politics futu...
5 0.01855 love iamonfire life learn man great will...
6 0.02446 news politics worldnews cbs obama state...
7 0.01233 entertainment fukushimaagain news chern...
8 0.0209 sports news local seahawks bowl super ga...
9 0.01302 make things trust fear didn't shit sense...
  
```

Topic Labels (Human)

```

0 Feel Good/Motivation
1 Books/Film
2 Local Headlines/Crime
3 Feel Good/Motivation
4 Iranian Film
5 Feel Good/Motivation
6 International Headlines
7 Nuclear Issues
8 Sports
9 Feel Good/Motivation
  
```



Results show that in January 2015, topics were loose, vague, and had both positive and negative affect.

Note that some topics were repeated, likely due to the fact that there were actually fewer than 10 total topics that the IRA chose to tweet about in that month.

Topics (Automated)

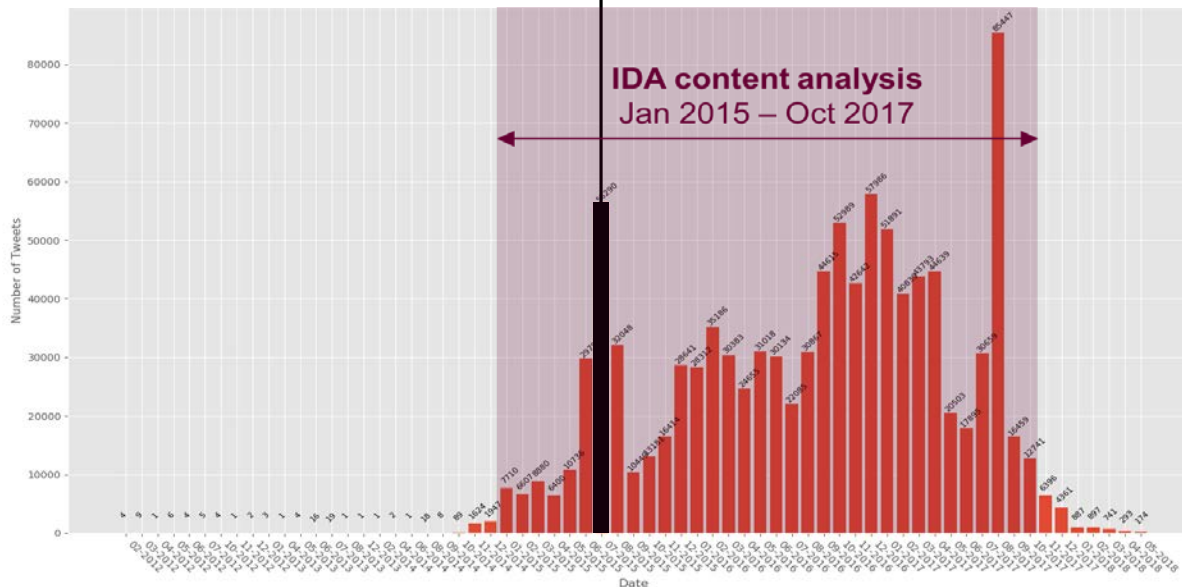
```

0 0.01507 workout exercise fitness good http://t...
1 0.02767 news world u.s sports deal iran obama b...
2 0.01754 exercise workout fat good don't lol run...
3 0.02011 weight lose loss fat diet don't eat lol ...
4 0.01764 workout exercise diet eat good today tom...
5 0.02082 workout good exercise today morning gre...
6 0.0284 news local sports man police cleveland ...
7 0.01846 workout morning good exercise today tom...
8 0.01862 workout fitness good gym today time gre...
9 0.04228 trump words politics don't fitness love ...
  
```

Topic Labels (Human)

```

0 Exercise/Nutrition
1 International Headlines
2 Exercise/Nutrition
3 Exercise/Nutrition
4 Exercise/Nutrition
5 Exercise/Nutrition
6 Local Headlines/Crime
7 Exercise/Nutrition
8 Exercise/Nutrition
9 Exercise + Donald Trump
  
```



In Summer 2015, topics were:

- Dominated by a single cultural issue (e.g., Exercise/Nutrition)
- Interspersed with news

Six months later, in July 2015, the pattern of topics had changed considerably. In Summer 2015, topics were dominated by a single cultural issue (e.g., Exercise/Nutrition) and were interspersed with news (e.g., Local Headlines/Crime).

Topics (Automated)

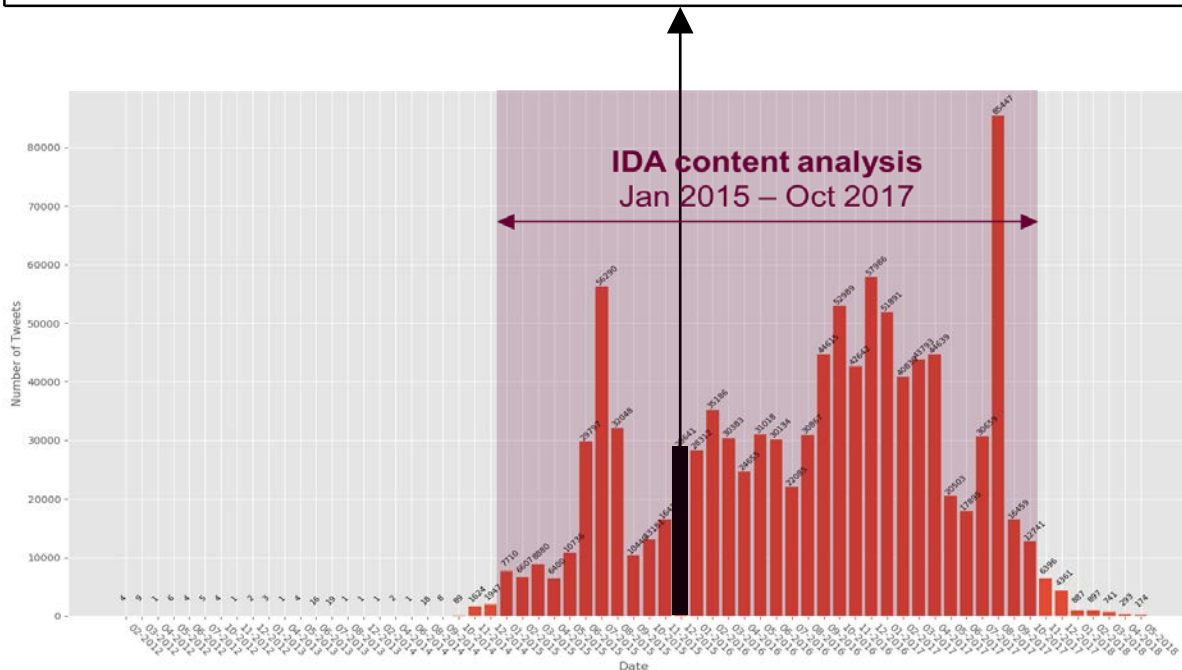
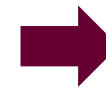
```

0 0.03043 news police man crime local shooting ch: ...
1 0.03199 news health business politics local tec: ...
2 0.01145 tcot texas pjnet today people usa obama ...
3 0.01934 christmas black people don't good day l: ...
4 0.02292 california prayers guns gun people tcot ...
5 0.02954 vegasgopdebate trump gopdebate politics ...
6 0.01285 news obama muslim san bernardino topnew: ...
7 0.01617 star words wars christmasin entertainme: ...
8 0.0251 news world syria u.s isis state islamic ...
9 0.02246 sports news orleans cleveland saints ch: ...
  
```

Topic Labels (Human)

```

0 Local Headlines/Crime
1 National Headlines
2 General Political Issues
3 Christmas
4 Guns
5 U.S. Presidential Primary
6 Terrorism
7 Christmas + Star Wars film
8 International Headlines
9 Sports
  
```



By late 2015, topics were:

- Tighter
- More specific
- More varied
- More polarizing

The pattern of topics changed again a few months later. Starting in late 2015, topics began to be tighter and more specific (e.g., Local vs. National vs. International Headlines), more varied (e.g., Christmas vs. Terrorism vs. Sports), and more polarizing (e.g., Guns, U.S. Presidential Primary). The tightness of the topics made it easier to manually assign a label to the topic.

Topics (Automated)

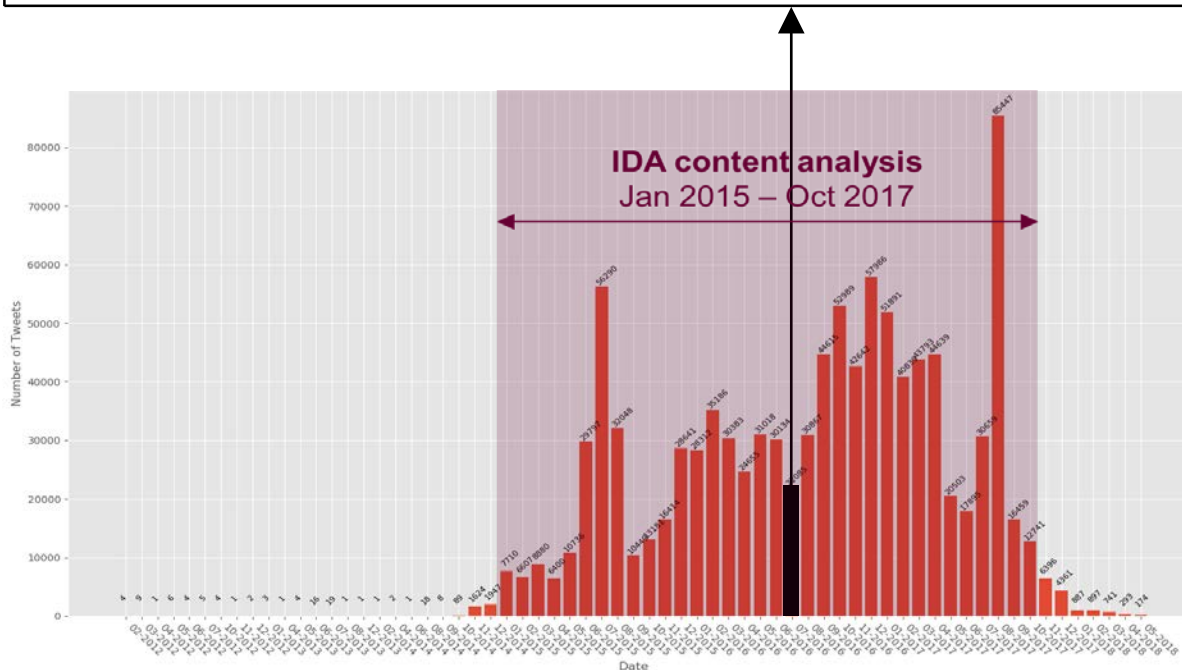
```

0 0.01824 words midnight timetravelerproblems wha' ...
1 0.02293 tech health news business u.s topnews e...
2 0.02744 politics trump clinton hillary conventi...
3 0.02664 ihatepokemongobecause ihate__because j...
4 0.0203 sports rio news game olympic breaking s...
5 0.02929 blackskinisnotacrime black blacklivesma'...
6 0.02933 news world attack u.s china south state...
7 0.0155 makeamovieblack black policeamovie that ...
8 0.02097 growingupwithobama hillary trump america...
9 0.02874 police news man shooting local topnews (...
  
```

Topic Labels (Human)

```

0 Hashtags
1 National Headlines
2 U.S. Dem/GOP Conventions
3 Hating Pokemon Go
4 Sports
5 Racial Issues
6 International Headlines
7 Racial Issues
8 U.S. Presidential Election
9 Local Headlines/Crime
  
```



The same pattern continued into mid 2016. Topics were still:

- Tight
- Specific
- Varied
- More polarizing

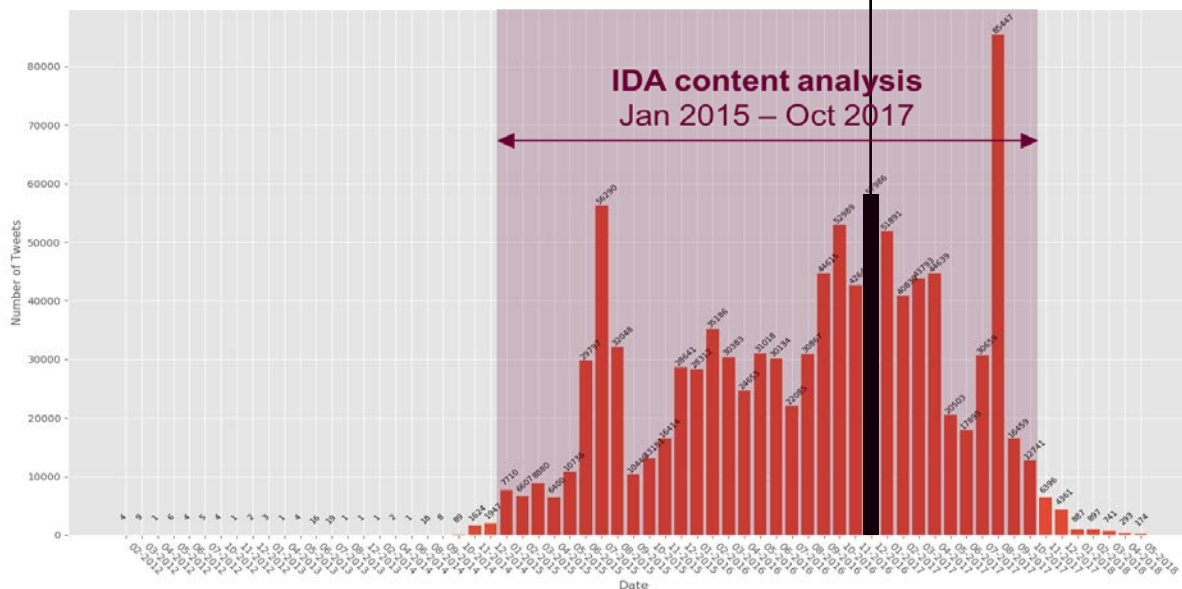
Jumping ahead to the following summer in July 2016, we see the same pattern of topics from late 2015: a variety of tight, easy-to-label topics that were more specific, more varied, and often more polarizing than in the first few months of our analysis (back in early 2015).

Topics (Automated)

0	0.0548	trump obama donald president state coll...
1	0.03572	christmas aleppo holiday city isis weatl...
2	0.04823	police man fire woman killed shot shoot...
3	0.02288	thingsyoucantignore alternativeacronymi...
4	0.03127	game win alternativeacronyminterpretati...
5	0.07697	black people amp white don't women god '...
6	0.03184	music nowplaying amp playing video jgtv...
7	0.08546	words christmas thingsyoucantignore lov...
8	0.03327	giftideasforpoliticians secondhandgifts ...
9	0.04789	trump news obama russia russian fake el...

Topic Labels (Human)

0	U.S. Presidential Election
1	International Headlines
2	Local Headlines/Crime
3	No topic
4	Sports
5	No topic
6	Music
7	Hashtags
8	Christmas + Hashtags
9	U.S. Presidential Election



The same pattern continued into late 2016. Topics were still:

- Tight
- Specific
- Varied
- More polarizing

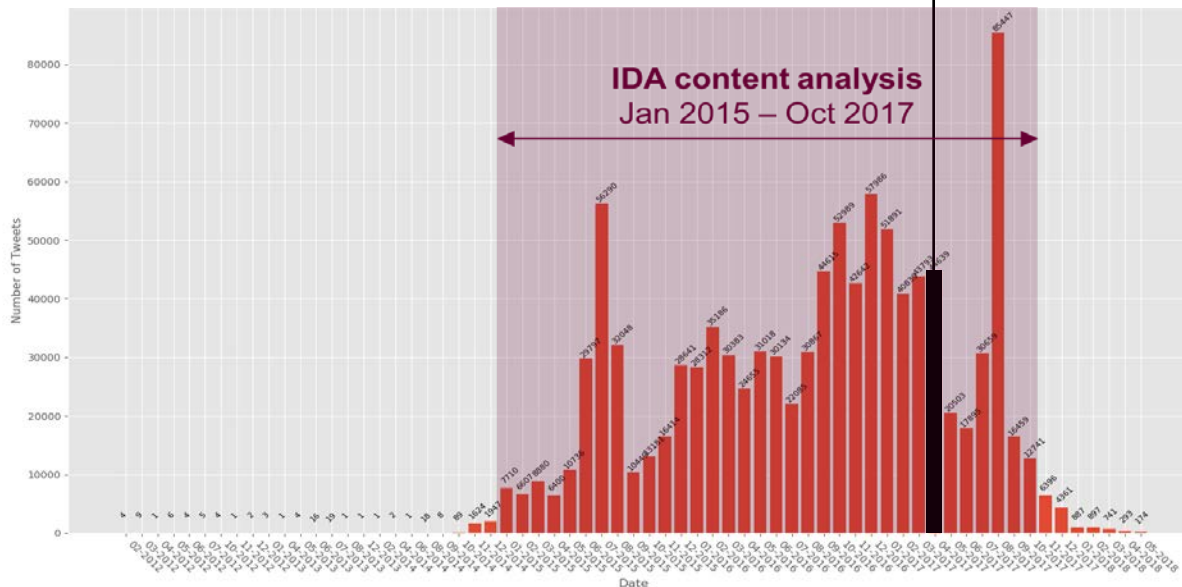
That same topic pattern continued into late 2016.

Topics (Automated)

0	0.05676	united april march san flight airlines : ...
1	0.04089	nowplaying amp video music listen playin...
2	0.0688	police man shooting killed woman found : ...
3	0.04272	game win nfl live draft series season s' ...
4	0.137	amp people don't black time good white : ...
5	0.0287	promotion soundcloud click music start j ...
6	0.08186	trump syria obama president russia news ...
7	0.04664	news bill muslim berkeley fox black isl ...
8	0.03457	words easterin midnight easter photos h ...
9	0.06591	trump tax court bill wall state city la ...

Topic Labels (Human)

0	National Headlines
1	Music
2	Local Headlines/Crime
3	Sports
4	No topic
5	Music
6	International Headlines
7	Terrorism
8	Hashtags
9	Immigration



The same pattern continued into early 2017. Topics were still:

- Tight
- Specific
- Varied
- More polarizing

The same topic pattern continued into early 2017.

Topics (Automated)

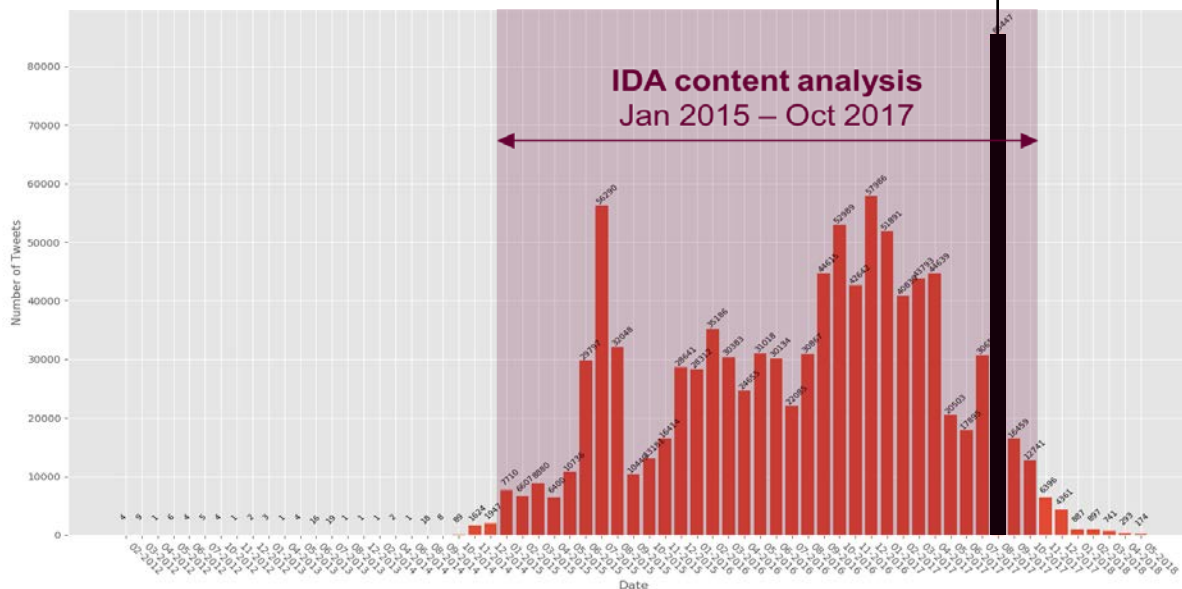
```

0 0.02199 trump breaking white charlottesville lil...
1 0.03452 hedgebz rewind https://t.co/mfbjiijvl ar...
2 0.02019 trump breaking news cnn fake media vide...
3 0.02295 trump charlottesville antifa video trum...
4 0.02256 mueller congress russia hillary trump m...
5 0.02391 maga trump retweet potus realdonaldtrum...
6 0.01707 enlist amp army https://t.co/mj patriot...
7 0.01649 trump jeff amb flake sanctuary city sch...
8 0.02075 korea trump north breaking obama kim mc...
9 0.08019 trump amp potus realdonaldtrump top peo...
  
```

Topic Labels (Human)

```

0 Charlottesville
1 No topic
2 Trump Presidency
3 Racial Issues
4 Mueller Investigation
5 Trump Presidency
6 General American Sentiment
7 Immigration
8 North Korean Nuclear Pgm.
9 Trump Presidency
  
```



August 2017 had by far the largest number of tweets. The same topic pattern continued into this month, now with mostly negative, polarizing topics (e.g., Charlottesville, Racial Issues, Mueller investigation).

Topics (Automated)

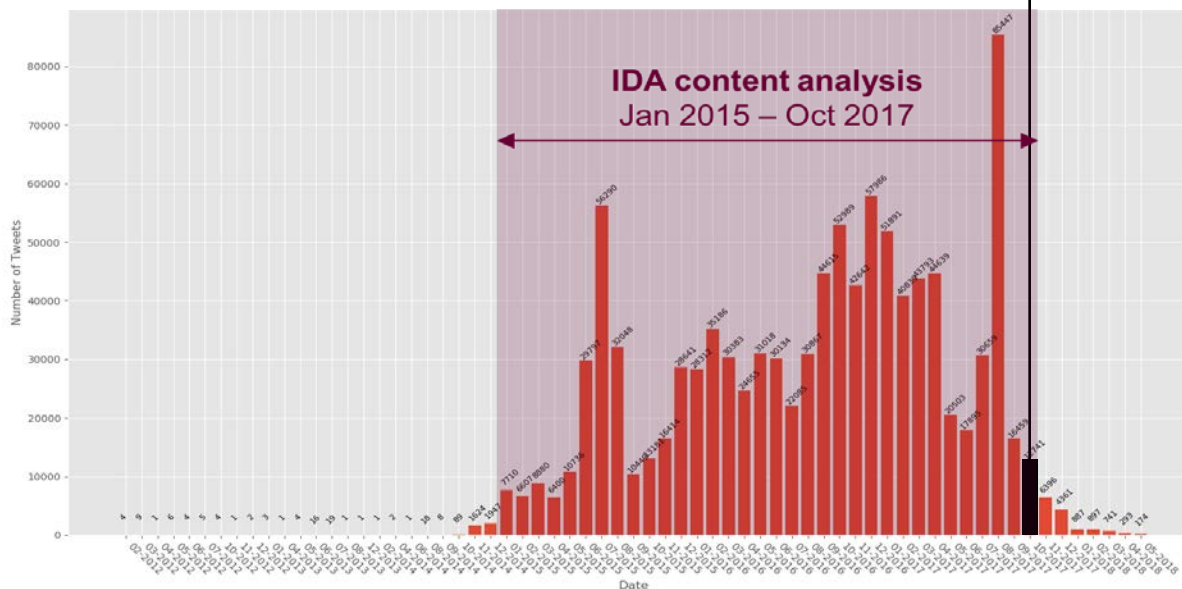
```

0 0.00313 breaking trump gun vegas shooting contro...
1 0.00373 trump breaking weinstein video traitor ...
2 0.00354 breaking hillary weinstein hollywood tr...
3 0.00359 trump san mayor puerto juan rico video l...
4 0.00315 trump claims hillary breaking deal obama...
5 0.00336 trump breaking video vegas media preside...
6 0.00415 weinstein harvey anthem nfl breaking pl...
7 0.00359 vegas massacre breaking video shooter l...
8 0.00321 breaking trump tweet destroys star virat...
9 0.00354 trump nfl breaking vegas lol kneelers w...
  
```

Topic Labels (Human)

```

0 Trump Presidency
1 General Political Issues
2 Me Too + Nat Anthem/Kaper.
3 Hurricane Maria/Puerto Rico
4 General Political Issues
5 Trump Presidency
6 Me Too + Nat Anthem/Kaper.
7 Vegas Shooting
8 Guns
9 Nat Anthem/Kapernick
  
```



In October 2017, topics were still:

- Tight
- Specific
- Negative affect only
- Mostly polarizing

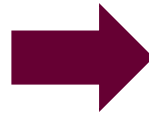
The final month our analysis, October 2017, showed the same pattern of topics. At this point, the topics are very tight, specific, negative, and polarizing. There are no longer any topics that imply positive affect and the topics are dominated by issues and events relevant to U.S. politics and partisanship (e.g., Trump Presidency, Me Too, National Anthem/Kaepernick, Hurricane Maria/Puerto Rico, Vegas Shooting, Guns).

January 2015

Topic Labels

- 0 Feel Good/Motivation
- 1 Books/Film
- 2 Local Headlines/Crime
- 3 Feel Good/Motivation
- 4 Iranian Film
- 5 Feel Good/Motivation
- 6 International Headlines
- 7 Nuclear Issues
- 8 Sports
- 9 Feel Good/Motivation

Late
2015



October 2017

Topic Labels

- 0 Trump Presidency
- 1 General Political Issues
- 2 Me Too + Nat Anthem/Kapernick
- 3 Hurricane Maria/Puerto Rico
- 4 General Political Issues
- 5 Trump Presidency
- 6 Me Too + Nat Anthem/Kapernick
- 7 Vegas Shooting
- 8 Guns
- 9 Nat Anthem/Kapernick

In January 2015, topics were:

- Loose
- Vague
- Positive and negative affect

By October 2017, topics were:

- Tight
- Specific
- Negative affect only
- Mostly polarizing

The IRA's English tweet topics grew **tighter**, more **specific**, more **negative**, and more **polarizing** over time, with the final pattern emerging in **late 2015**.

We now show a side-by-side comparison of the results from the first and last months of our content analysis:

- In January 2015, topics were loose, vague, and had both positive and negative affect. For example, topics like Books/Film, Sports, and Feel Good/Motivation dominated.
- In contrast, by October 2017, tight, specific, negative, and polarizing topics like Me Too, Colin Kapernick's protests of the National Anthem, hurricane destruction in Puerto Rico, and the Las Vegas shooting were the most prominent topics.

To summarize, the IRA's English tweet topics grew tighter, more specific, more negative, and more polarizing over time, with the final pattern emerging in late 2015. This pattern remained fairly constant for the next two years, at least through the end of our period of analysis (October 2017).

This evolution in the pattern of topics over time informs our hypothesis about what the IRA's intent may have been: First, establish a cover and gain followers. The IRA experimented with at least two early approaches— (1) the loose, vague, and both positive and negative topics of early 2015, as well as (2) a single dominating topic interspersed with news, like in mid 2015. Referring back to our earlier plot of IRA followings and followers over time, we see that the average number of followers (light grey) per active IRA handle did not increase much in this time period. However, the IRA began experimenting with a third approach— (3) tight, specific, negative, and polarizing topics in late 2015. Interestingly enough, our followings/followers plot showed a large average increase in followers per active IRA handle shortly after that, beginning in early 2016. That is, the IRA may have learned that tight, specific, more negative, and more polarizing topics led to followers, allowing them to settle on their winning strategy for influencing the largest numbers of English-speaking users in the United States.

Legend

■ Other

■ News

■ Entertainment/Culture

■ U.S. Politics

■ Polarizing Issue

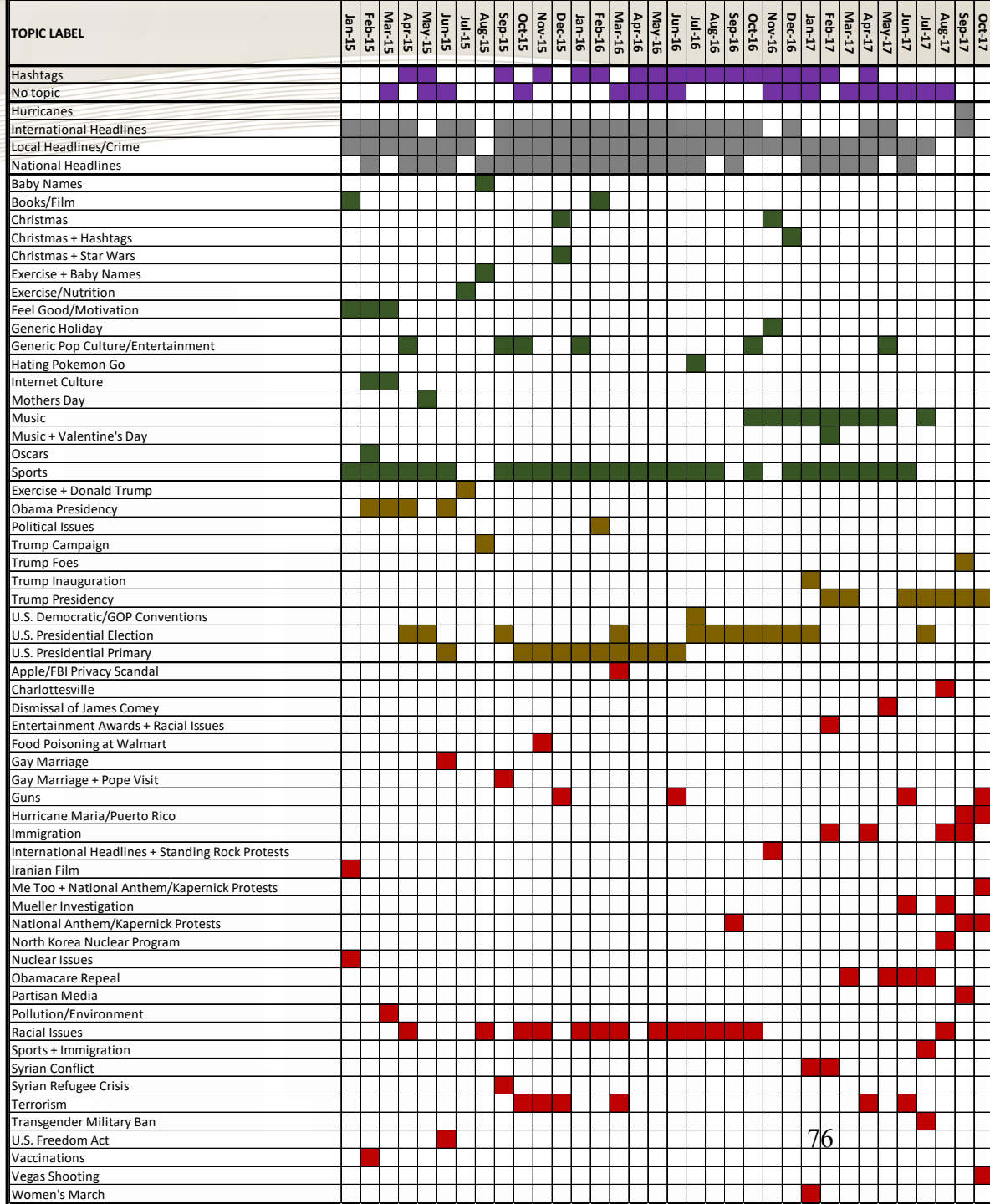
Furthermore...

Some of the IRA's English tweet topics were **short-lived**.

(e.g., Hating Pokemon Go, Trump Inauguration, Syrian Conflict)

Other topics were **long-lasting**.

(e.g., Headlines, Music, Sports, U.S. Presidential Election / Primary, Racial Issues)



This matrix summarizes all topics that appeared in the English training set tweets over our period of analysis, from January 2015 through October 2017. Each row corresponds to a topic, while each column correspond to a month. A square is filled with a color if that topic was represented in that month and is left blank otherwise. We grouped the topics into five different categories:

- (Grey) *News*: Topics taken from news headlines, including local crime reports, national headlines, and international news stories.
- (Green) *Entertainment/Culture*: Topics related to music, sports, holidays, films, and so forth.
- (Yellow) *U.S. Politics*: Topics describing American political elections, presidencies, debates, primaries, and so forth, without describing any specific political issue.
- (Red) *Polarizing issue*: Topics with two distinct sides for debate, often with strong right- vs. left-wing perspectives. Most of these topics were relevant to the United States.
- (Purple) *Other*: Topics consisting of random hashtags only and/or very few non-hashtag words, such that we could not assign a label to the topic.

Early in our period of analysis (left), there were more Entertainment/Culture topics (green) than Polarizing Issues topics (red). The opposite trend was present towards the end of our period of analysis (right)– more Polarizing Issues than Entertainment/Culture (more red than green).

Furthermore, many of the IRA's English tweet topics were short-lived. Naturally, many of these short-lived topics referred to events with limited time spans, such as the Trump Inauguration. Furthermore, other short-lived topics referred to brief fads, such as Hating Pokemon Go. However, other topics were short-lived even though related events continued over long time spans in the real world, such as the Syrian Conflict. It may be that the IRA did tweet about topics like the Syrian Conflict over many months, but at a volume low enough such that the LDA statistical model merged these tweets into the broader International Headlines topic. That is, perhaps the IRA only tweeted about the Syrian Conflict in a large enough volume to form its own topic in only two months, January and February 2017.

Other topics were long-lasting, such as Headlines, Music, Sports, the U.S. Presidential Election/Primary, and Racial Issues. We hypothesize that the IRA used the Headlines, Music, and Sports topics to attract a large American followership and maintain cover, while the Racial Issues topic was used to sow discord among the American followers, which was then directed towards political events and issues such as the U.S. Presidential Election/Primary.

This page is intentionally left blank.

“That use of social media— that weaponization of social media— began as early as 2012. Was significantly up and running by 2013. Was full bore by 2014 – 2015, and it wasn’t until late 2016 and even into 2017 that the Intelligence Community really got a handle of what was going on. **And had we identified it much earlier, say in 2012 – 2013**, that the Russians were going to conduct this kind of attack on the United States— had the President known that— **we could’ve had more options** than he ended up having in the Summer of 2016.”

– Michael Morell, Acting Director of the CIA, 2012 – 2013
(Podcast Interview, Summer 2019)

Yes, but...

Bearing these results in mind, we now consider the broader implications.

This issue has been considered by current and former members of the Intelligence Community. For example, Michael Morell, former Acting Director of the CIA in 2012 – 2013, discussed the IRA's social media activities on a publicly available podcast in Summer 2019.²³ He stated, "That use of social media— that weaponization of social media, began as early as 2012. Was significantly up and running by 2013. Was full bore by 2014 – 2015, and it wasn't until late 2016 and even into 2017 that the Intelligence Community really got a handle of what was going on. *And had we identified it much earlier, say in 2012 – 2013, that the Russians were going to conduct this kind of attack on the United States— had the President known that— we could've had more options* than he ended up having in the Summer of 2016." [emphasis added]

This statement is certainly true— in retrospect, had the DoD and Intelligence Community identified the IRA's social media activities sooner, the United States could have had more options to prevent any potential impact on American interests. However...

²³ Bharara, Preet. 2019. CIA Spies & Gadgets (with Michael Morell), minutes 56:03-56:41. *Stay Tuned with Preet Bharara*.
<https://podcasts.apple.com/us/podcast/stay-tuned-with-preet/id1265845136?i=1000445846835>

- **Early identification** of the IRA's use of Twitter against the U.S. would have been **difficult**, since:
 - The IRA's English tweet topics evolved over time
 - Their final pattern did not emerge until late 2015
- **The U.S. government must expect that our adversaries' social media activity will evolve over time**
- Efficient **processing pipelines** are needed for analyses of time-evolving social media activity:
 - IDA has set up a pipeline for semi-autonomous analysis of tweets → **we can now very quickly process a very large number of very small documents**
 - IDA could improve its pipeline even further...

Early identification of the IRA's use of Twitter and other social media platforms against the United States would have been difficult, since the IRA's English tweet topics evolved over time and their final pattern did not emerge until late 2015, only one year before the 2016 American Presidential election.

Thus the United States government must expect that our adversaries' social media activity will continue to evolve over time. We must be able to anticipate and respond to our adversaries' adaptation.

To do that, efficient processing pipelines are needed for analyses of time-evolving social media activity:

- For example, IDA has set up a pipeline for semi-autonomous analysis of tweets. We can now quickly process a large number of very small documents (tweets).
- Such a pipeline could be improved even further...

- Apply extensions to the LDA statistical model to **automatically track the evolution of tweet topics over time** (i.e., evolve hidden states via Markov process)
- Extend analysis to **other languages** (e.g., German, Russian, etc)
- Create **dashboard** to let analyst drill down into the tweets:



For these suspected IRA handles:

- Show me all tweets posted in German in the last month...
- Of these, show me all that were related to the topic label "*Syrischer Konflikt*" (Syrian Conflict)...
- This tweet looks interesting. Show me all tweets like this one from Sept 2015 (when Germany opened its border to Syrian refugees)...
- Of these, show me all tweets that included the hashtag #OpIceISIS...
- Etc...

The next steps of our analysis involve three different paths:

1. From a **science and technology** perspective, there are opportunities for improving the underlying statistical model used to extract topics from each month's corpus of tweets. In this analysis, a human was required to manually perform two simultaneous tasks:
 - Assign a 1-4 word label to each topic word list *and*
 - Recognize if/when a topic in one month was the same as a topic in another month.

Both of these tasks required the human labeler's time and attention. However, a more detailed statistical model could perform the second task automatically, allowing the human to focus solely on the first task. The more detailed statistical model could evolve the hidden states (topics) via a Markov process.²⁴ **Automated tracking of topics over time could significantly speed up analysis, allowing for a much larger corpus of tweets (millions, rather than merely thousands, of tweets per month) and a more high-resolution sampling in time (one corpus per week, rather than one corpus per month).**

2. From a **geopolitical** perspective, we could repeat our same analysis (or the extended statistical analysis described above) on the IRA's non-English tweets. Our results showed that although a large majority (approximately 71%) of the IRA training set tweets were in English, a large percentage (approximately 21%) were in Russian, as well as German (approximately 3%) and Ukrainian (approximately 1%). The same statistical techniques and process pipelines from this analysis could be used to extract topics from the Russian, German, and Ukrainian tweets, as well. Human labelers with language fluency in Russian, German, and Ukrainian could then assess those topics and assign a 1-4 word label to each topic word list. To perform this task well, the human labelers would need both foreign language expertise plus cultural knowledge of Russian, German, Ukrainian, and neighboring countries' affairs, in order to best understand what different words in the topics mean to Twitter users in those countries. (For example, the human labeler in our⁸⁵ analysis

had an innate understanding of American life and therefore knew the social and cultural impact of the phrases “Black Lives Matter” and “Oscar So White” in English.) **Extending our analysis into foreign languages would provide insight into how the IRA’s tactics and strategy differed between the United States and other countries.**

3. From a **human factors** and **operations research** perspective, we could create a dashboard to allow an analyst to drill down into the tweets. For instance, one could imagine a situation in which an intelligence analyst was asked to monitor all tweets posted by several dozen accounts all likely to be owned by an adversary to the United States. The analyst could ask himself or herself:

“For these suspected adversary handles:

- *Show me all tweets posted in German in the last month...*
- *Of these, show me all that were related to the topic label “Syrischer Konflikt” (“Syrian Conflict”)...*
- *This tweet looks interesting. Show me all tweets like this one (based on the same topics) from September 2015 (when Germany opened its borders to Syrian refugees)...*
- *Of these, show me all tweets that included the hashtag #OpIcISIS...*
- *And so forth.”*

A graphical user interface (GUI) would not take the place of the human intelligence analyst. Instead, it would function as a decision support tool to help analysts answer their own questions about a corpus of tweets, allowing them to incorporate their own creativity and geopolitical savviness into the process (two characteristics for which humans will continue to outperform machines for the foreseeable future.)

IDA has the scientific/technical, foreign language/geopolitical, and human factors/operations research skill sets to address each of these next steps.

- ²⁴ Blei, David M. and Lafferty, John D. 2006 Jun. Dynamic Topic Models. *Proceedings of the 23rd International Conference on Machine Learning*. <https://dl.acm.org/doi/10.1145/1143844.1143859>



Giant Military Cats

421 Tweets



Following

Giant Military Cats

@GiantCat9

Just giant cats with military hardware.

 Joined September 2019

0 Following **42K** Followers

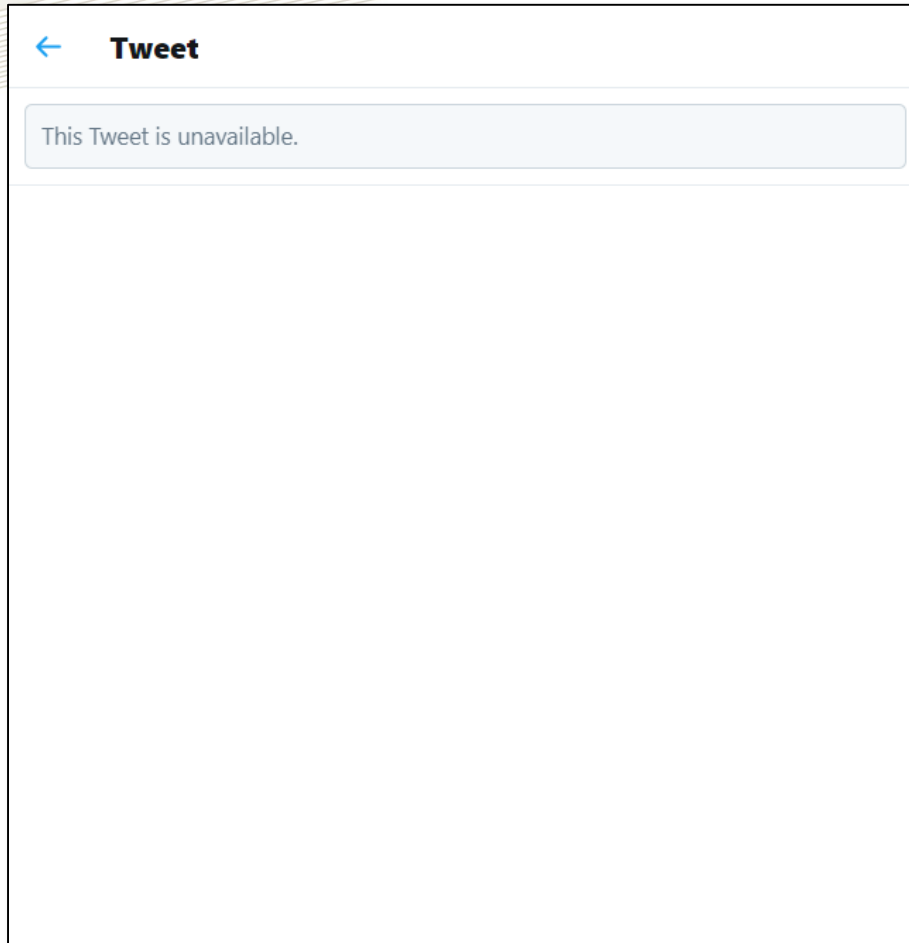
Please contact IDA with any questions:

Dr. Shelley Cazares
Institute for Defense Analyses
4850 Mark Center Drive
Alexandria, VA 22311
703 845 6792
scazares@ida.org

²⁵ Giant Military Cats. Twitter. <https://twitter.com/giantcat9>

Backups

The following slides present additional information that may be relevant to follow-on discussions.



Twitter has now deleted this tweet – and all other tweets posted by @TodayInSyria

An early slide showed a mock-up example of how an IRA tweet would have appeared on Twitter. Twitter has now removed this tweet, along with all other tweets posted by this same handle. Therefore, if one were to log onto Twitter and navigate to the URL for this tweet, one would simply see a blank screen that said, “This Tweet is unavailable.”²⁶

²⁶ <https://twitter.com/3899481526/status/832645688471101440>

IDA | Other Example IRA Tweets

@TheFoundingSon

Teach your kid gun safety after Wikileaks tells you all about cheese pizza [#PizzaGate](#)
<https://t.co/WGubutlb4r>

6:58 PM • Nov 16, 2016

39255 Followers

@SouthLoneStar

RT if you think that [#NationalPizzaDay](#) is the perfect day for Jeff Sessions to expose [#PizzaGate](#) pedophile ring.

8:44 PM • Feb 9, 2017

33530 Followers

@TheFoundingSon

Just a reminder that James Alefantis is not your everyday pizza shop owner, and [#PizzaGate](#) is entirely real
<https://t.co/Z0mAcQrmfa>

1:44 PM • May 31, 2017

39255 Followers

@SouthLoneStar

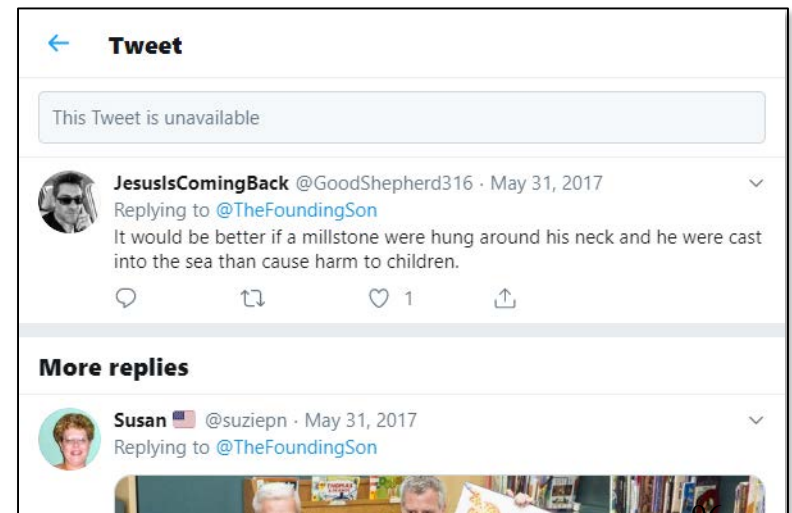
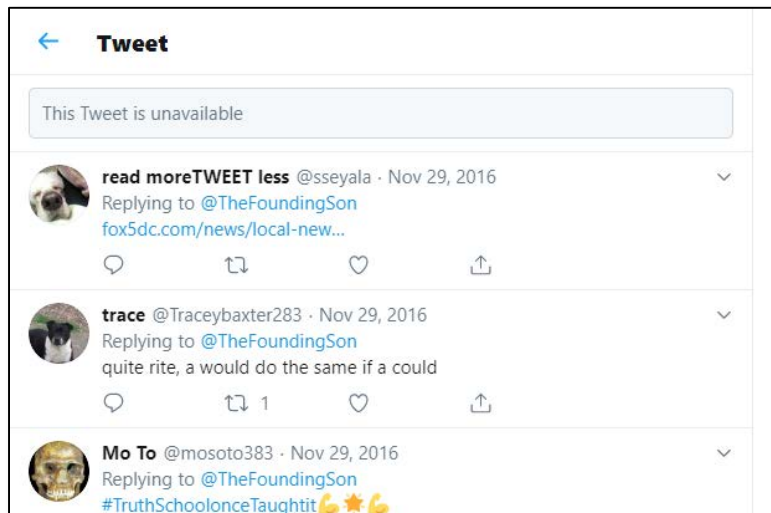
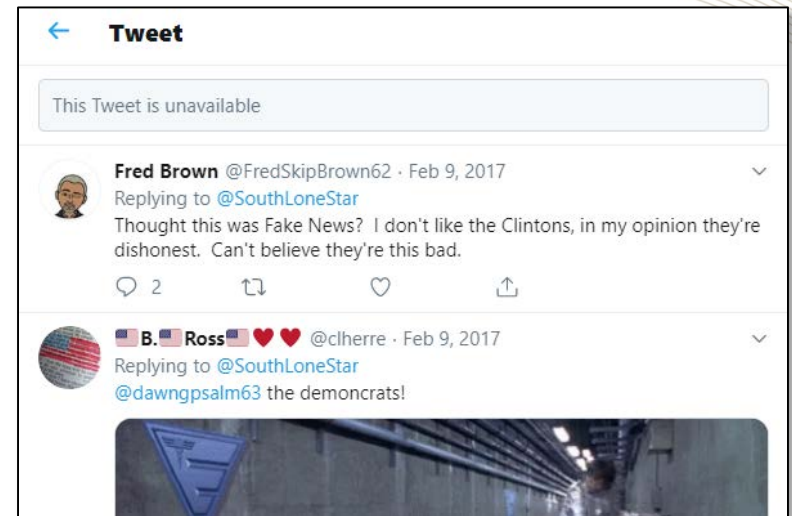
Sean Spicer just said in his press conference that children are being sold in the United States. [#Pizzagate](#) confirmed by White House.

1:29 AM • Feb 24, 2017

38941 Followers

An early slide showed a mock-up example of one IRA tweet. This slide shows four additional mock-up examples of four additional IRA tweets. The two tweets on the left were posted by one handle, @TheFoundingSon. The two tweets on the right were posted by another handle, @SouthLoneStar.

IDA | Other Example IRA Tweets (now deleted from Twitter)



The previous backup slide showed four mock-up examples of four IRA tweets. All four tweets have now been deleted from Twitter, along with all other tweets posted by the same handles. If one were to log onto Twitter and navigate to any of these four tweets, one would see the statement, “This Tweet is unavailable.”^{27,28,29,30} However, comments made by other, non-IRA handles in response to each tweet can still be seen.

²⁷ <https://twitter.com/4355797094/status/834938300905435136>

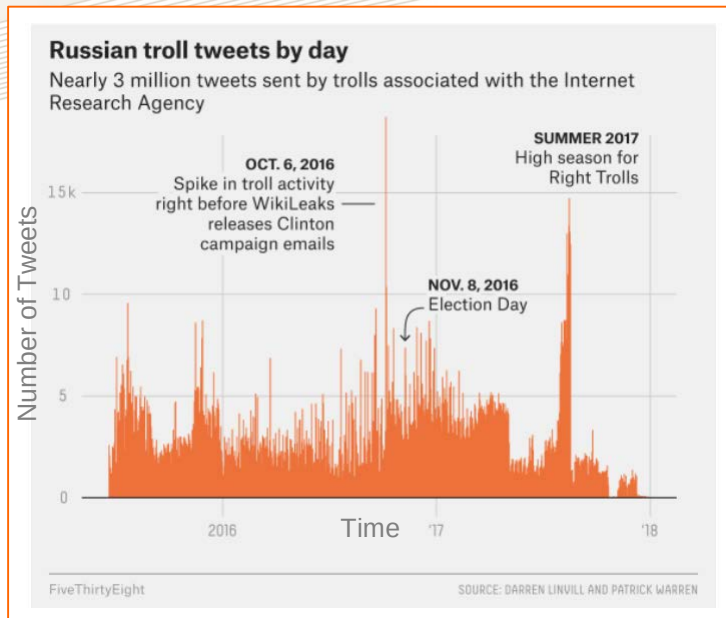
²⁸ <https://twitter.com/4355797094/status/829793128760672256>

²⁹ <https://twitter.com/4508630900/status/803674493273128960>

³⁰ <https://twitter.com/4508630900/status/869912349418770433>

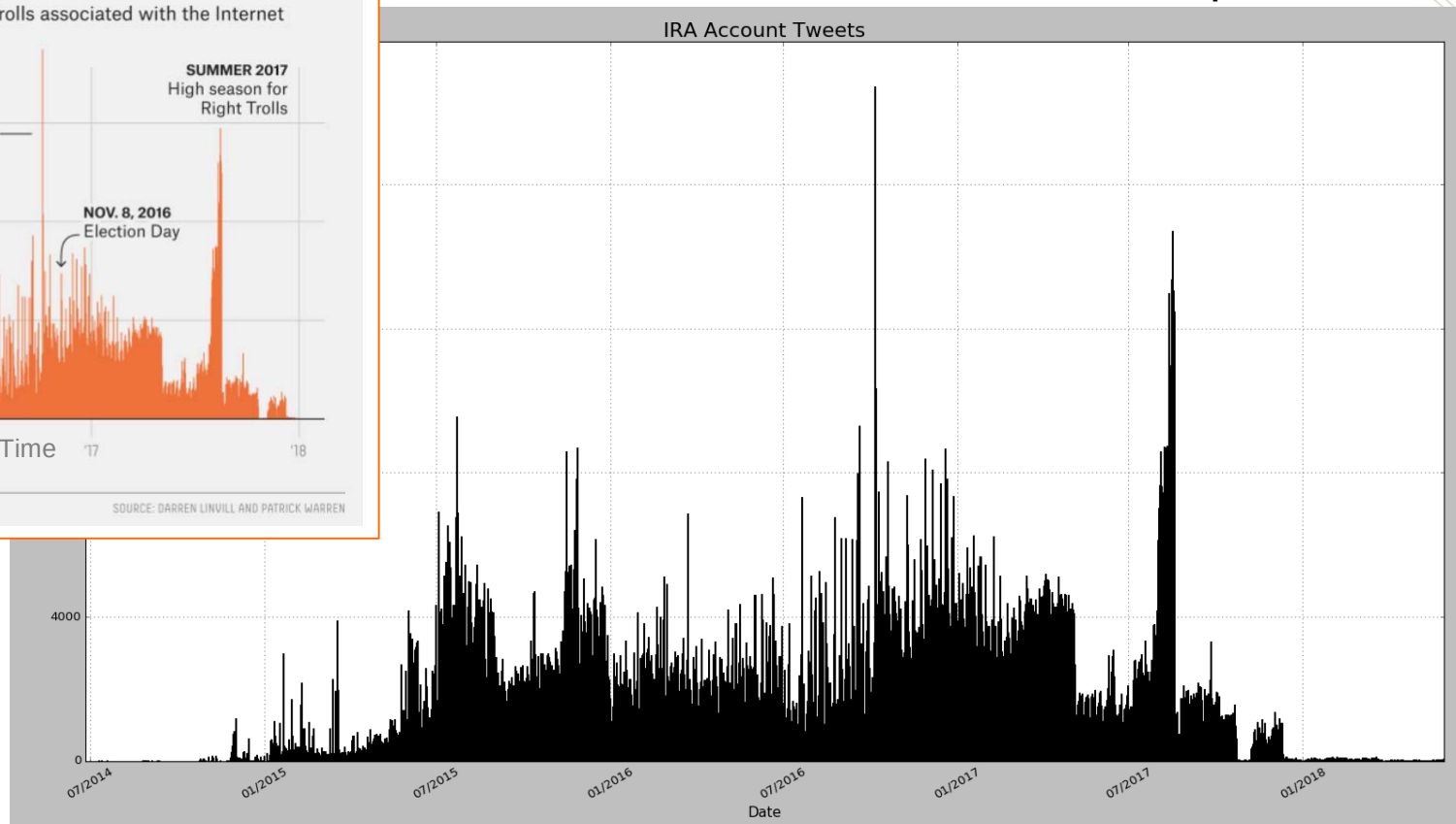
IDA | IDA duplicated the *FiveThirtyEight* results

From *FiveThirtyEight* Article:



IDA replicated results:

Histograms
(binned per day)
of total IRA
tweets over time

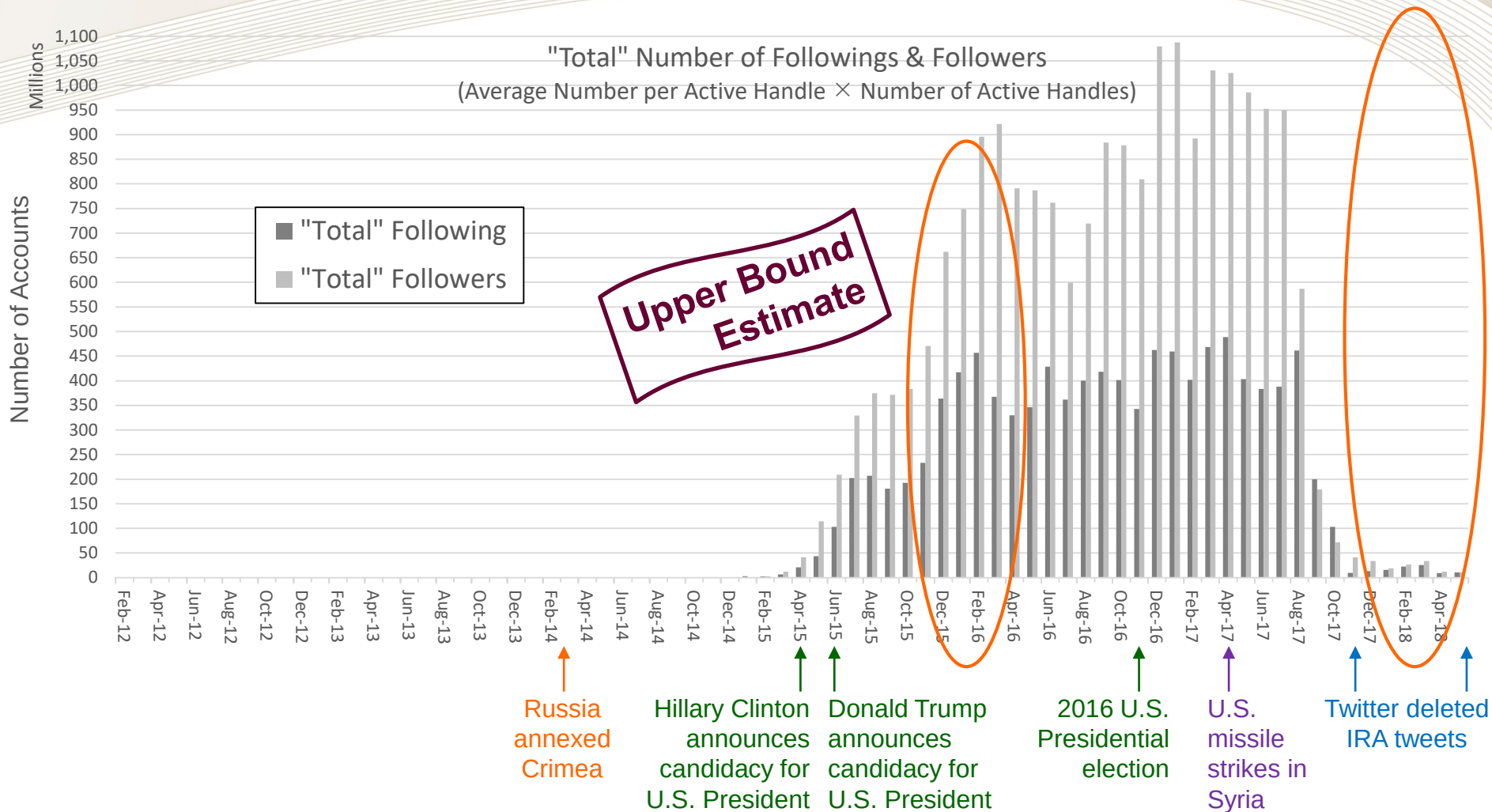


IDA identified the same trends as *FiveThirtyEight* in Number of IRA Tweets vs. Time, with peaks aligning to major events. 98

The polling aggregation website *FiveThirtyEight* published Clemson's dataset of IRA tweets. They also published a document explaining the content of the dataset, as well as an initial plot of the number of IRA tweets over time (orange), noting interesting events from the real world that correlated with spikes in the histogram.³¹ We quickly replicated this plot (black) early in our analysis.

³¹ Roeder, Oliver. 2018 Jun 2013. Why We're Sharing 3 Million Russian Troll Tweets. *FiveThirtyEight*.
<https://fivethirtyeight.com/features/why-were-sharing-3-million-Russian-troll-tweets/>

IDA investigated IRA followers & followings over time



Limitations of available metadata → Difficulties estimating total IRA audience
(e.g., may double-count the same followers of two IRA accounts)

An earlier slide plotted the number of active IRA handles (accounts) over time, while its following slide plotted the average number of followings and followers over time, per active IRA handle. Multiplying these two metrics together on a month-by-month basis, we create an upper bound estimate of the total number of IRA followings (dark grey) and followers (light grey) over time. We note that this is merely an *upper bound* estimate, due to the possibility that multiple IRA handles could have shared the same followings and followers.

IDA | Indiana University's bot detector

Botometer®

An OSoMe project (bot•o•meter)



Botometer (formerly BotOrNot) checks the activity of a Twitter account and gives it a score based on how likely the account is to be a bot. Higher scores are more bot-like.

Use of this service requires Twitter authentication and permissions. ([Why?](#))

If something's not working or you have questions, please contact us only after reading the [FAQ](#).

Botometer is a joint project of the Network Science Institute ([IUNI](#)) and the Center for Complex Networks and Systems Research ([CNETS](#)) at Indiana University.

▼  @GeorgeMichael  0.5 / 5  ×

English-specific features

Content: 0.8

Sentiment: 0.9

Language-independent features

Friend: 0.7

Network: 0.8

Temporal: 1.3

User: 0.1

Bot score based on



All features: 0.5

Language-independent: 0.3

Tweet language: en

Complete Automation Probability: 0%



Profile



Tweet



Details



Feedback

← Not Bot-like

▼  @GMichaelBot 4.2 / 5  ×

English-specific features

Content: 3.3

Sentiment: 2.1

Language-independent features

Friend: 2.1

Network: 3.0

Temporal: 3.1

User: 3.0

Bot score based on



All features: 4.2

Language-independent: 3.2

Tweet language: en

Complete Automation Probability: 59%



Profile



Tweet



Details



Feedback

← Bot-like

Other organizations have used supervised machine learning techniques to create troll or bot detectors in social media:

- A troll is a person who creates discord in social media by purposely posting an inflammatory message to “rile up” his or her audience.
- A bot is an autonomous system that posts messages in social media in a manner that appears to be human-like.

This slide shows an example of Indiana University’s Botometer,³² applied to two different twitter handles: @GeorgeMichael (blue, top) and @GMichaelBot (orange, bottom). As described by its developers, “Botometer (formerly BotOrNot) checks the activity of a Twitter account and gives it a score based on how likely the account is to be a bot. Higher scores are more bot-like.”³³

- (Blue, Top) @GeorgeMichael is the personal account of the pop singer George Michael. This handle earns a Botometer score of only 0.5 out of 5, indicating a low likelihood of being a bot.
- (Orange, Bottom) @GMichaelBot is a well-known bot designed to automatically retweet any tweet about George Michael. This handle earns a Botometer score of 4.2 out of 5, indicating a high likelihood of being a bot.

Botometer is an example of a supervised machine learning system designed to automatically detect objects of interest (i.e., bots). To train this system, the Botometer developers required a training set of data (tweets), where each tweet was labeled by whether or not it was truly posted by a bot or a human.³⁴ Labeling such a dataset with a high degree of certainty can be a very resource-intensive process.

Many other research groups are attempting to do the same thing for Russian trolls— develop a supervised machine learning system to automatically detect Russian trolls. To train such a system, developers require a training set of tweets where each tweet was labeled by whether or not it was truly posted by a Russian troll or a regular Twitter user. Labeling this kind of dataset with a high degree of certainty could also be a very resource-intensive process.

In contrast, IDA’s analysis of the IRA Twitter activity used an unssupervised machine learning system designed to assist a human analyst in summarizing and understanding a set of tweets. To train our system, we also required a training set of tweets, taken from the Clemson dataset. However, our analysis did not require ground truth labels for the tweets themselves; instead the underlying LDA statistical model learned the best way to organize the tweets in order to associate them with different topics.

Our statistical model had no way to know what any particular topic *meant*, however. A human was then needed to label these topics. Fortunately, the individual tweets did not have to be labeled, since their associations with each topic had already been quantified by the statistical model. Labeling topics is a much less resource-intensive process than labeling the tweets themselves— while there were over 1.4 million English tweets in our training set, there were only $34 \text{ months} \times 10 \text{ topics/month} = 340$ topics over our period of analysis (i.e., from January 2015 through October 2017).

³² Botometer. <https://botometer.iuni.iu.edu/#/>

³³ *ibid.*

³⁴ *ibid.*

This page is intentionally left blank.

- Growing scholarly attention to trolls on social media includes:
 - Analyzing trolls' manipulation of political discourse and/or **quantifying their impact** on Presidential elections
 - Using graphical analysis or social theory to understand troll behavior and strategy
 - Using machine learning for detection of bots, trolls, or incivility
- A smaller group of researchers have published on Clemson's IRA dataset:
 - Kim *et al.* (2019) note: "...**temporal information is often ignored or disregarded** in analysis...the focus has been on tweet volume and hashtag frequency...rather than how roles and strategies change over time."

Growing scholarly attention has been paid in recent years to social media trolls. These efforts include:

- Analyzing trolls' manipulation of political discourse and/or quantifying their impact on Presidential elections,³⁵
- Using graphical analysis³⁶ or social theory³⁷ to understand troll behavior and strategy, and
- Using machine learning for detection of trolls, bots, or incivility.^{38,39,40}

A smaller group of researchers have published on Clemson's dataset. As described by Kim et al. (2019), "... temporal information is often ignored or disregarded in analysis... the focus has been on tweet volume and hashtag frequency... rather than how roles and strategies change over time."⁴¹

Therefore, IDA's approach to semi-automated tracking of the IRA's tweet topic *patterns over time* could fill a void in the social media research community.

³⁵ Badawy, Adam; Emilio Ferrara; and Kristina Lerman. 2018. Analyzing the Digital Traces of Political Manipulation: The 2016 Russian Interference Twitter Campaign. *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. Barcelona. 258-265.

³⁶ Hindman, Matthew and Barash, Vlad. 2018. *Disinformation, 'Fake News' and Influence Campaigns on Twitter*. Miami: Knight Foundation.

³⁷ Kim, Dongwoo; Graham, Timothy; Wan, Zimin; and RizoIU, Marian-Andrei. 2019. Analyzing user identity via time-sensitive semantic edit distance (t-SED): A case study of Russian trolls on Twitter. *Journal of Computational Social Science* 2:331-351.

³⁸ Im, Jane; Chandrasekharan, Eshwar; Sargent, Jackson; Lighthammer, Paige; Denby, Taylor; Bhargava, Ankit; Hemphill, Libby; Jurgens, David; and Gilbert, Eric. 2019. Still Out There: Modeling and Identifying Russian Troll Accounts on Twitter. *arXiv:1901.11162 [cs.SI]*.

³⁹ Mihaylov, Todor and Nakov, Preslav. 2016. Hunting for Troll Comments in News Community Forums. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Berlin: Association for Computational Linguistics. 399-405.

⁴⁰ Davis, Clayton A.; Varol, Onur; Ferrara, Emilio; Flammini, Alessandro; and Menczer, Filippo. 2016. BotOrNot: A System to Evaluate Social Bots. *WWW '16 Companion: Proceedings of the 25th International Conference Companion on World Wide*. Montreal. 273-274.

⁴¹ Kim, Dongwoo; Graham, Timothy; Wan, Zimin; and RizoIU, Marian-Andrei. 2019. Tracking the Digital Traces of Russian Trolls: Distinguishing the Roles and Strategy of Trolls On Twitter. *arXiv:1901.05228 [cs.CY]*.

IDA | LDA Equations

Probability distribution of corpus $\mathcal{C}$:

$$p(\mathcal{C}) = \left(\prod_{i=1}^K p(\beta_i | \eta) \right) \left(\prod_{d=1}^D p(\theta_d | \alpha) \left[\prod_{n=1}^{N_d} p(z_{dn} | \theta_d) p(w_{dn} | z_{dn}, \beta) \right] \right)$$

Fit all V words from all D docs in corpus $\mathcal{C}$ to β_i
(i^{th} row of word-topic matrix):

$$\beta = \begin{bmatrix} \vdots & \ddots & \vdots \\ \vdots & \ddots & \vdots \\ \vdots & \ddots & \vdots \end{bmatrix}$$

Words in corpus: 1 to V
Topics in corpus: 1 to K

(ij^{th} element estimates probability that j^{th} word is associated with i^{th} topic)

Fit all N_d words in d^{th} doc to θ_d
(d^{th} doc's topic mixing vector):

$$\theta_d = \begin{bmatrix} \vdots & \dots & \vdots \end{bmatrix}$$

Topics in corpus: 1 to K

(i^{th} element estimates probability that i^{th} topic is associated with d^{th} doc)

Look up probability of z_{dn} (topic that generated n^{th} word in d^{th} doc) from θ_d (d^{th} doc's topic mixing vector)

Look up probability of w_{dn} (n^{th} word in d^{th} doc) from β (word-topic matrix)

V = Number of words in corpus (regardless of what doc they came from)

K = Number of topics in corpus (regardless of what doc they came from)

β = Word-topic matrix (one matrix per corpus)

D = Number of docs in corpus

$\theta_d = d^{th}$ doc's topic mixing vector (one vector per doc)

N_d = Number of words in d^{th} doc

z_{dn} = Topic that generated n^{th} word in d^{th} doc (used to look up correct row of β and correct element of θ_d)

$w_{dn} = n^{th}$ word in d^{th} doc (used to look up correct column of β)

Latent Dirichlet Allocation (LDA) is a well-known statistical model used in unsupervised machine learning problems since 2003.^{42,43} It is a probabilistic generative model used for topic modeling of text-based documents.

LDA differs from a standard mixture model:

- With standard mixture models, all words of a document (i.e., tweet) are assumed to have been generated from a single topic. That is, each document (i.e., tweet) is associated with one and only one topic.
- In contrast, with LDA, that assumption does not apply— each word of a document (i.e., tweet) may have been generated by a different topic. That is, each tweet may be associated with multiple topics.

Like many statistical models, LDA treats some variables as observables and others as hidden (i.e., latent) variables:

- Observables are variables that the user can see and know for certain, such as each word of each tweet. The n^{th} word of the d^{th} tweet is called w_{dn} .
- In contrast, hidden (i.e., latent) variables are those that the user cannot see or know for certain. Instead, the user must attempt to *estimate* what these variables are. In LDA, the hidden variables are: z_{dn} , θ_d , α , β , and η , as we discuss below.

This slide writes out the probability distribution for a corpus of documents C (i.e., for a month of tweets). This is the joint probability distribution all variables written above: w_{dn} , z_{dn} , θ_d , α , β , and η :

- β is a matrix that keeps track of how strongly each unique word in the corpus is associated with each topic. Each row of β refers to one of the K topics in the corpus, while each column refers to one of the V unique words in the corpus. The i,j^{th} element of β estimates the probability of the j^{th} unique word in the i^{th} topic. Since β is a hidden variable, we cannot see or know it for certain. Therefore, we must estimate the values of the elements of matrix β , given all words in all documents (observables) and η (another hidden variable, drawn from the Dirichlet distribution).
- θ_d is a vector that keeps track of how strongly each document in the corpus is associated with each topic. Each element of θ_d refers to one of the K topics in the corpus. The i^{th} element of θ_d estimates the probability of the i^{th} topic in the d^{th} document. θ_d is also a hidden variable; therefore we must estimate it, as well. We do this given all words in the d^{th} document (observables) and α (another hidden variable, also drawn from the Dirichlet distribution).
- z_{dn} is the topic from which the n^{th} word in the d^{th} document is drawn. We can look up the probability of z_{dn} from θ_d .
- w_{dn} is the probability of the n^{th} word in the d^{th} document. We can use z_{dn} to look up the probability of w_{dn} from β .

Several different statistical techniques can be used to fit the elements of β and θ_d to the data, given α and η , respectively. We used the software package MALLET.⁴⁴

⁴² Blei, David M.; Ng, Andrew Y.; and Jordan, Michael I. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning*, 3: 993-1022.

<http://jmlr.org/papers/volume3/blei03a/blei03a.pdf>

⁴³ Blei, David. 2019 Nov 2. Topic Models. *VideoLectures.net*. http://videolectures.net/mlss09uk_blei_tm/

⁴⁴ MALLET Machine Learning for Language Toolkit. <http://mallet.cs.umass.edu/>

This page is intentionally left blank.



Lenalyo

@lenalyo

Автопробег выпускников Академии ЦРУ The rally CIA Academy graduates #Dallas

<https://t.co/xdGLQsAhoi>

Azerbaijan

2:31 PM • Jul 8, 2016

118 Following • 119 Followers



} English or Russian?

So far, IDA's content analysis has focused on the IRA's English tweets, since, according to the Clemson metadata language tags, most of the tweets were posted in English. However, before performing our content analysis, we first explored Clemson's metadata language tags in more detail.

This slide depicts another example of an IRA tweet. It includes text in two languages:

- Russian: “Автопробег выпускников Академии ЦРУ”
- English: “The rally CIA Academy graduates #Dallas”

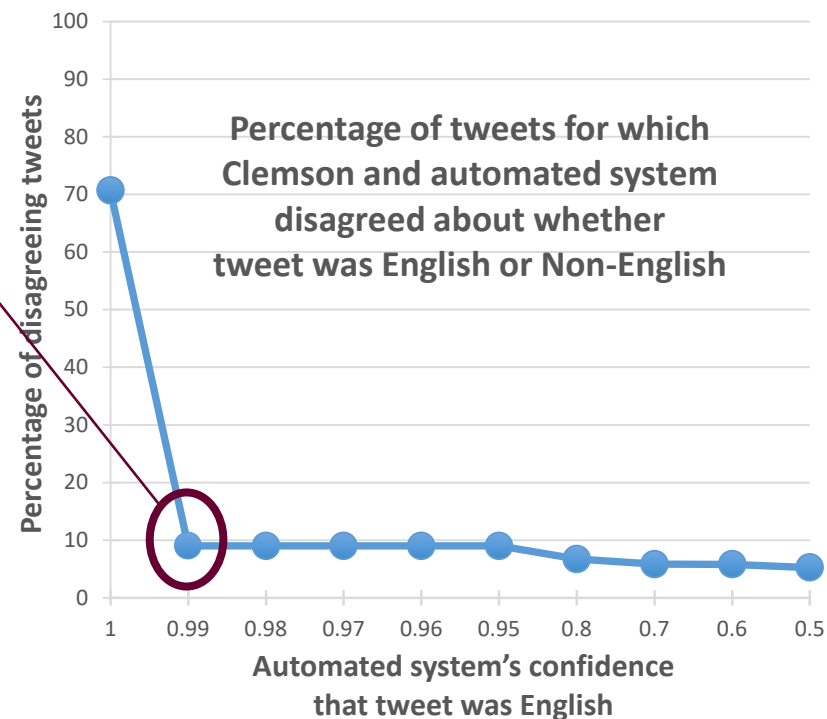
Clemson tagged this particular tweet as Russian. However, the methods they used to assign this tag is unclear. Furthermore, the English text in this tweet could be useful for an English analysis, as well. To investigate this issue further, we used automated techniques to classify each individual tweet by language and then compared our automated classifications to Clemson's metadata language tags, as we will explain on the next slide.

Language identification is not always straightforward for tweets

Clemson

		English	Non-English
Automated System	English	908,752 tweets (61.7%)	0 tweets (0%)
	Non-English	91,445 tweets (6.2%)	7,210 tweets (0.5%)
		46,667 tweets (3.2%)	419,029 tweets (28.4%)

(1,431,103 total tweets in training set)



Although Clemson's language labels weren't perfect, they were "good enough" to support the next step of our content analysis.

We used the Python software library *langdetect*⁴⁵ to automatically classify the language of each individual tweet. This library automatically identifies a set of candidate languages for a document (i.e., a tweet) and then estimates the confidence that the document is written in each of those candidate languages. For the purposes of our analysis, all tweets with an English confidence greater than or equal to threshold were classified as English, while all tweets with an English confidence less than threshold were classified as non-English. Furthermore, all tweets for which English was *not* identified as a candidate language at all were also classified as non-English. (Thus there were two different ways in which a tweet could be classified as non-English.)

Right plot: We explored what value should be used for the confidence threshold in our analysis. On the right, we plot the percentage of training set tweets for which the Clemson language tags and our automated system (based on *langdetect*) disagreed about whether the tweet was English or non-English, as the confidence threshold was varied. In this plot, disagreements include those tweets which Clemson tagged as English but the automated system classified as non-English, and vice versa. The confidence thresholds indicate how difficult it was for the automated system to classify a tweet as English:

- A confidence threshold of 1.00 was a difficult bar to surpass, such that not very many tweets could be classified as English by the automated system, leading to a large (approximately 70%) disagreement with Clemson's English tags.
- In contrast, a confidence threshold of 0.50 was a much lower bar to surpass, such that many more tweets could be classified as English by the automated system, leading to a small (approximately 5%) disagreement with Clemson's English tags.

The shape of this plot indicates that the percent disagreement between Clemson's English tags and automated English classifications is not very sensitive to the confidence threshold, for all confidence thresholds of 0.99 and below. Therefore, we chose a confidence threshold of 0.99 for our analysis.

Left table: We then investigated the types of disagreements between the automated system and the Clemson English tags, with the confidence threshold held constant at 0.99. On the left, we show a 2×2 confusion matrix comparing the English vs. non-English classifications of our automated system vs. Clemson:

- The green boxes along the diagonal indicate the number of training set tweets for which the two methods agreed. That is, of the 1,431,103 total tweets in the training set:
 - (Green, Top Left) 908,752 (61.7%) were classified as English by both Clemson and the automated system.
 - (Green, Bottom Right) 426,239 (28.9%) were classified as non-English by both methods. We broke out the two different ways in which tweets could be classified as non-English by the automated system, leading to *two* green boxes in the lower, right corner of the confusion matrix: Of these 426,239 tweets for which the two methods agreed were non-English:
 - 419,029 (28.4%) were classified as non-English by the automated system because the automated system did not identify English as a candidate language at all.
 - 7,210 (0.5%) were classified as non-English by the automated system because although English was identified as a candidate language, the English confidence did not surpass the 0.99 threshold.
- The red boxes indicate the number of training set tweets for which the two methods *disagreed* (i.e., those tweets forming the circled point in the plot to the right). Of the 1,431,103 training set tweets:
 - (Red, Top Right) Zero were classified as English by the automated system but non-English by Clemson.
 - (Red, Bottom Left) Conversely, 138,112 (9.4%) were classified as non-English by the automated system and English by Clemson. Of these:

- 46,667 (3.2%) were classified as non-English by the automated system because the automated system did not identify English as a candidate language at all.
- In contrast, while the automated system identified English as a candidate language for the remaining 91,445 (6.2%) tweets, the English confidence did not surpass the 0.99 threshold.

The confusion matrix shows that, overall, the automated system was more cautious than Clemson when classifying a training set tweet as English based on a confidence threshold of 0.99. This trend persisted for all other potential confidence threshold values, as well.

To explore the causes for the disagreements between the automated system and Clemson, we randomly sampled 300 training set tweets from each of the two red boxes in the lower, left corner of the confusion matrix.⁴⁶ Visual assessment of these tweets shows that disagreements between the Clemson tags and the automated classifications generally occurred for one of two reasons:

- The tweet was a combination of English and another language, such as the example tweet shown on the previous slide. In that example, Clemson tagged the tweet as Russian, even though it contained both Russian and English text. In most cases, though, Clemson tagged similar tweets as English.
- The tweet contained people's names, cities, or other proper nouns that appeared to be English, even though other parts of the tweet were in another language.

In most cases, Clemson's tags appeared to be most correct, per our subjective, qualitative judgment. That is, when Clemson and the automated system disagreed about whether a tweet was English or non-English, we usually sided with Clemson.

We then considered which English vs. non-English classification method should be used for the remainder of our analysis, in order to isolate the English tweets in the training set for subsequent topic modeling:

- On one hand, the automated system could scale easily to an even larger set of new tweets, and could therefore be used in prospective intelligence analysis.
- On the other hand, the more cautious behavior of the automated system classified many tweets as non-English, even when those tweets contained some English text that could still prove useful in topic modeling.
- Furthermore, although Clemson's tags would not be available to process new tweets in the future, visual assessment showed them to be most correct for this particular dataset.

Therefore, we chose to use Clemson's tags for the remainder of our analysis on this particular dataset. Such tags will not be available for future tweets, though, in which case we will need to resort to the automated method, instead.

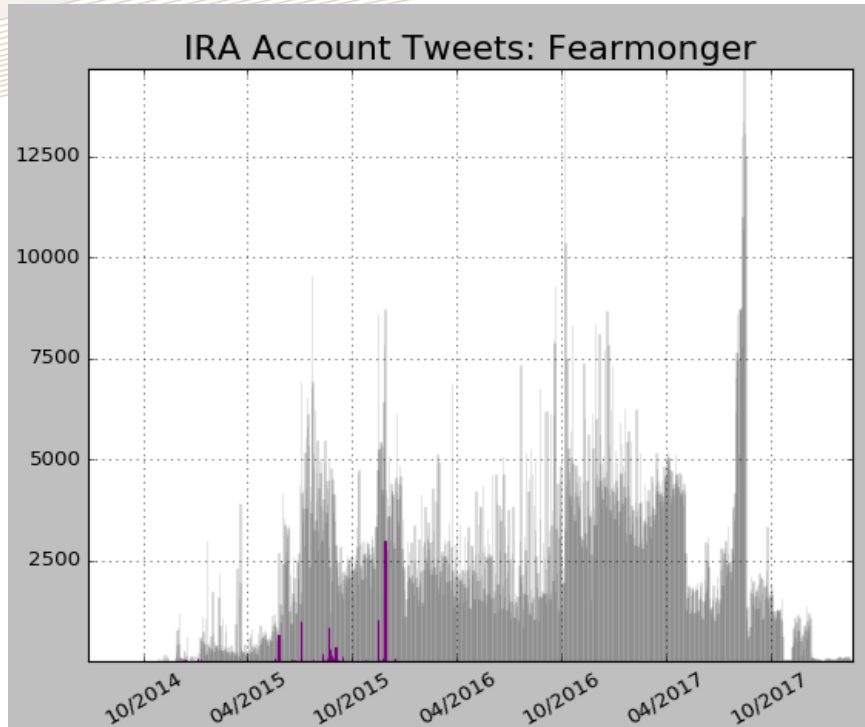
⁴⁵ langdetect 1.0.7. <https://pypi.org/project/langdetect/>

⁴⁶ We visually assessed 600 training set tweets: 300 which Clemson tagged as English but the automated system classified as non-English because it did not identify English as a candidate language at all, plus another 300 training set tweets which Clemson tagged as English but the automated system classified as non-English because although it identified English as a candidate language, the English confidence did not surpass the 0.99 threshold. There were zero tweets which Clemson tagged as non-English but the automated system classified as English.

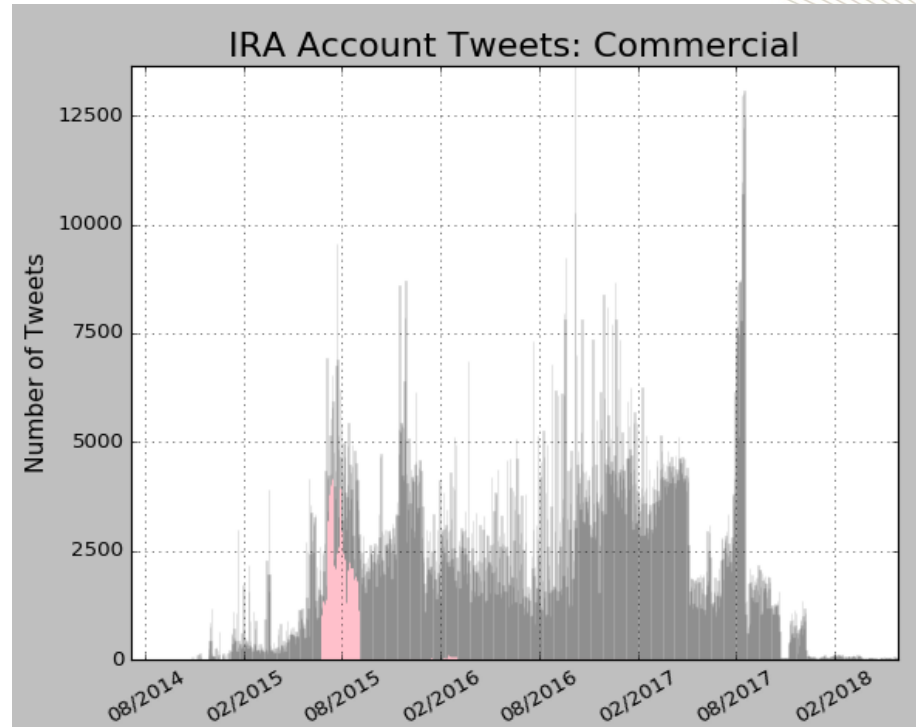
This page is intentionally left blank.

IDA | IDA reviewed Clemson's manual labels of IRA tweets (1 of 3)

The color plots the number of IRA tweets per day with the specified label. These were plotted against all IRA tweets (in gray).



“**Fearmongers**, relatively rare in the data set, spread news about a fake crisis, such as salmonella-contaminated turkeys around Thanksgiving, or the toxic chemical fumes described at the beginning of the New York Times Magazine article about the Internet Research Agency.”



Commercial handles “engaged in commercial activity (marketed exercise and diet related activities, marketed payday loans, and marketed essay writing services).”

Fearmonger and Commercial IRA tweets never picked up steam

As discussed in a previous slide, Clemson provided one final piece of metadata— a six-category scheme to summarize the overall theme of each IRA handle (i.e., left troll, right troll, fearmonger, commercial, hashtag gamer, and news feed). This slide defines two of these categories, fearmongers and commercial, and plots some trends. (The next two slides show definitions and trends for the four other categories.) Each plot shows the total number of tweets over time, binned by day and shown in grey. The grey plot is identical between the two figures. However, superimposed in color on top are the number of tweets posted by handles that Clemson categorized as fearmonger (purple, left) and commercial (pink, right):

- (Purple, Left) *Fearmongers*: As described by FiveThirtyEight, “Fearmongers, relatively rare in the data set, spread news about a fake crisis, such as salmonella-contaminated turkeys around Thanksgiving, or the toxic chemical fumes described at the beginning of the New York Times Magazine article about the Internet Research Agency.”⁴⁷ Our plots show that few tweets were posted by fearmonger handles in 2015, and almost none were posted by fearmonger handles after late 2015, the point in time at which the IRA appeared to settle on its winning tweet topic pattern.
- (Pink, Right) *Commercial*: As the Clemson researchers explain, commercial handles “engaged in commercial activity (marketed exercise and diet related activities, marketed payday loans, and marketed essay writing services).”⁴⁸ Commercial handles were active in mid 2015 only, the point in time at which the IRA was still employing an early tweet topic pattern in which they focused on one dominating topic (e.g., Exercise/Fitness).

⁴⁷ Roeder, Oliver. 2018 Jun 2013. Why We’re Sharing 3 Million Russian Troll Tweets. *FiveThirtyEight*.

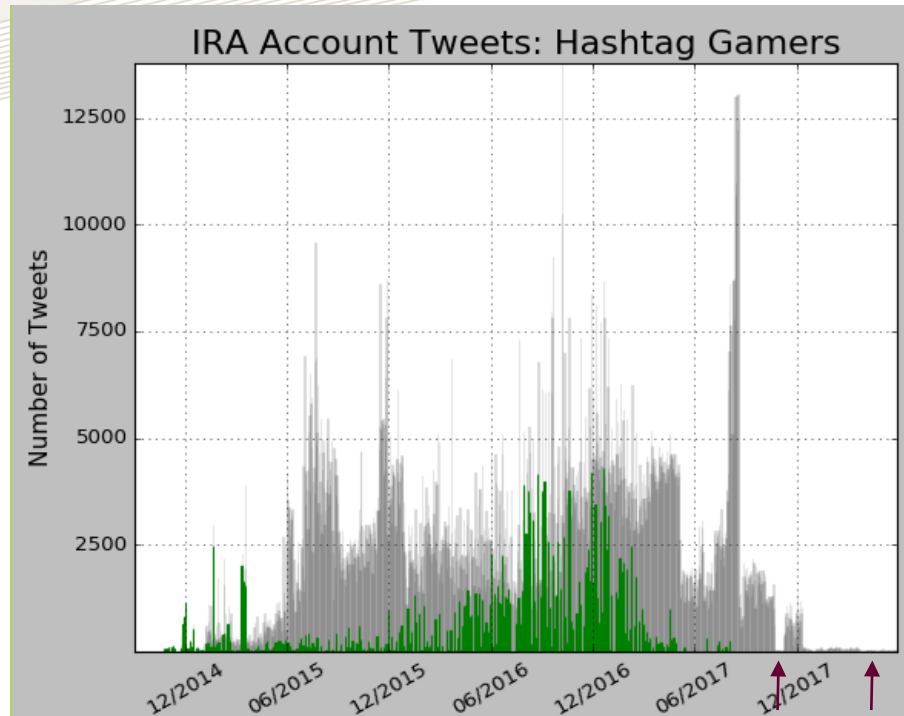
<https://fivethirtyeight.com/features/why-were-sharing-3-million-Russian-troll-tweets/>

⁴⁸ Linvill, Darren L. and Warren, Patrick L. 2018. *Troll Factories: The Internet Research Agency and State-Sponsored Agenda Building*.

http://pwarren.people.clemson.edu/Linvill_Warren_TrollFactory.pdf

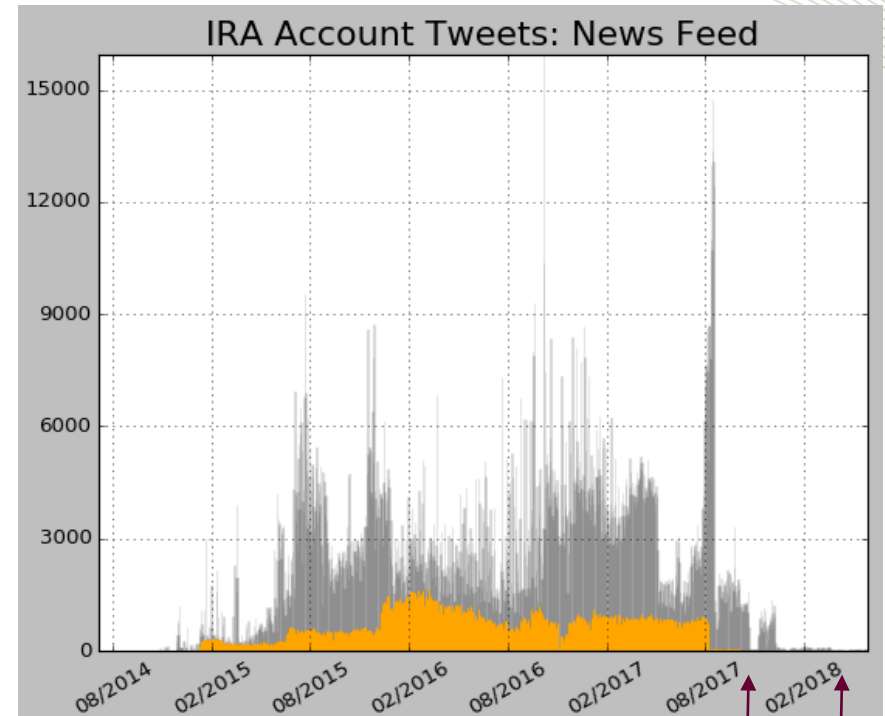
IDA | IDA reviewed Clemson's manual labels of IRA tweets (2 of 3)

The color plots the number of IRA tweets per day with the specified label. These were plotted against all IRA tweets (in gray).



Twitter deleted
IRA handles

“**Hashtag Gamers** specialize in playing hashtag games (e.g., #LessInterestingBooks might give rise to the tweet “Waldo’s Right Here”); many of their tweets are harmless wordplay in the spirit of the games, but some are socially divisive, in the style of Right Trolls or Left Trolls.”



Twitter deleted
IRA handles

“They [**News Feeds**] present themselves as local news aggregators, with names such as @OnlineMemphis and @TodayPittsburgh, and the news they link to is typically legitimate.”

News Feed and Hashtag Gamer IRA tweets lost steam even before Twitter deletion

This slide defines and shows trends for two more of Clemson's six handle categories: hashtag gamers (green, left) and news feeds (yellow, right):

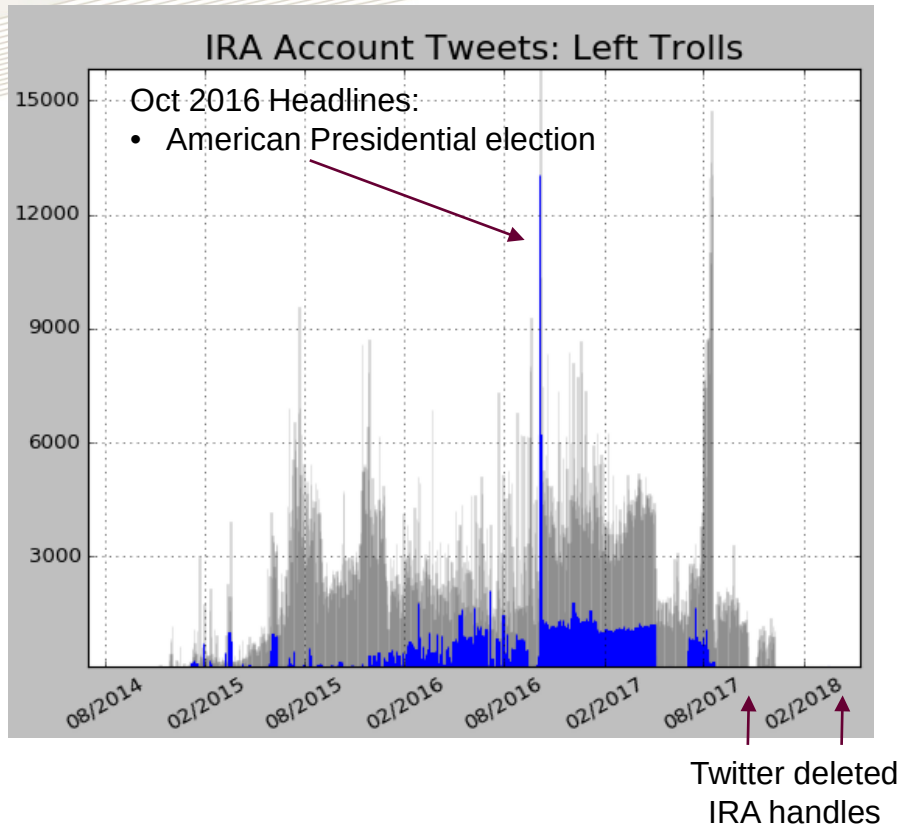
- (Green, Left) *Hashtag gamer*: As *FiveThirtyEight* explains, "Hashtag Gamers specialize in playing hashtag games (e.g., #LessInterestingBooks might give rise to the tweet "Waldo's Right Here"); many of their tweets are harmless wordplay in the spirit of the games, but some are socially divisive, in the style of Right Trolls or Left Trolls."⁴⁹ Hashtag gamer handles were most active in 2016. This corroborates the results of our topic modeling analysis, for which our final topics vs. month matrix shows that the hashtag topic persisted through 2016.
- (Yellow, Right) *News feed*: "They present themselves as local news aggregators, with names such as @OnlineMemphis and @TodayPittsburgh, and the news they link to is typically legitimate."⁵⁰ News feed handles consistently tweeted over the course of our entire period of analysis. This trend also corroborates the results of our topic modeling analysis, for which our final matrix shows the News topics persisting throughout the entire period of analysis.

⁴⁹ Roeder, Oliver. 2018 Jun 2013. Why We're Sharing 3 Million Russian Troll Tweets. *FiveThirtyEight*.
<https://fivethirtyeight.com/features/why-were-sharing-3-million-Russian-troll-tweets/>

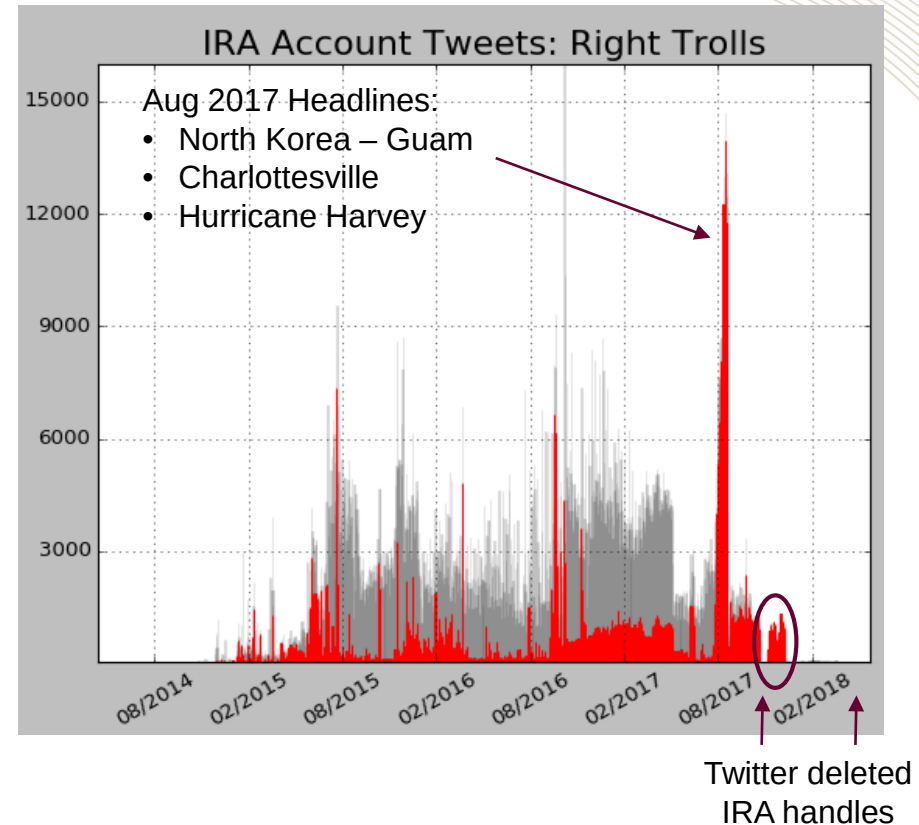
⁵⁰ *ibid.*

IDA | IDA reviewed Clemson's manual labels of IRA tweets (3 of 3)

The color plots the number of IRA tweets per day with the specified label. These were plotted against all IRA tweets (in gray).



“**Left Trolls** often adopt the personae of Black Lives Matter activists, typically expressing support for Bernie Sanders and derision for Hillary Clinton, along with clearly trying to divide the Democratic Party and lower voter turnout.”



“**Right Trolls** behave like bread-and-butter MAGA Americans, only all they do is talk about politics all day long.”

Right Troll tweets regained steam after Twitter deletion, but Left Troll tweets did not

This slide explores Clemson's final two handle categories: left and right trolls, both of which refer to U.S. political partisan perspectives:

- (Blue, Left) “*Left Trolls* often adopt the personae of Black Lives Matter activists, typically expressing support for Bernie Sanders and derision for Hillary Clinton, along with clearly trying to divide the Democratic Party and lower voter turnout.”⁵¹ The left troll handles were most active in October 2016, immediately before the 2016 U.S. Presidential election.
- (Red, Right) “*Right Trolls* behave like bread-and-butter MAGA Americans, only all they do is talk about politics all day long.”⁵² The right troll handles peaked almost a year later in August 2017. Looking back towards that period in time, we remember news headlines such as the U.S.-North Korean standoff over Guam, the white supremacist protests in Charlottesville, and Hurricane Harvey ravaging Houston, TX.

The red circle on the right indicates that the right troll (red) handles quickly regained steam immediately after the first Twitter suspension in November 2017, while the left troll (blue) handles did not.

⁵¹ Roeder, Oliver. 2018 Jun 2013. Why We're Sharing 3 Million Russian Troll Tweets. *FiveThirtyEight*.
<https://fivethirtyeight.com/features/why-were-sharing-3-million-Russian-troll-tweets/>

⁵² *ibid.*

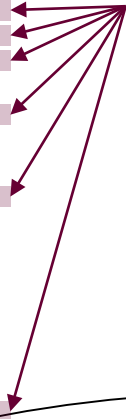
Russian, by Region

German, by Region

English, by Region

Region	2012	2013	2014	2015	2016	2017	2018	Total
Afghanistan	0	0	0	3	271	187	0	461
Austria	0	0	0	0	0	0	0	0
Azerbaijan	0	0	242	691	4228	1750	0	6911
Belarus	0	0	0	145	130	174	0	449
Canada	0	0	0	600	0	0	0	600
Czech Republic	0	0	0	0	2	0	0	2
Denmark	0	0	0	0	0	1	0	1
Egypt	0	0	0	22	37	0	0	59
Estonia	0	0	0	0	0	0	0	0
Finland	0	0	0	3	0	4	1	8
France	0	0	0	1	39	27	0	67
Germany	0	0	0	13	1015	257	0	1285
Greece	0	0	0	5	2	5	0	12
Hong Kong	0	0	0	0	1	0	0	1
India	0	0	0	0	0	0	0	0
Iran, Islamic Republic of	0	0	0	0	1	19	0	20
Iraq	0	0	0	142	27	258	0	427
Israel	0	0	0	245	164	0	0	409
Italy	0	0	0	0	0	1125	0	1125
Japan	0	0	0	0	0	0	0	0
Latvia	0	0	0	3	0	0	0	3
Malaysia	0	0	0	2	1	1	0	4

Same
Bin



Former military intelligence analysts at IDA provided context, binning the IRA tweets from Russia's geopolitical perspective

In order to further understand potential trends and patterns within the IRA metadata, tweets were broken down by language and geographical region and compared to one another over time. The IRA tweeted in 42 languages, supposedly from 21 regions. However, some languages and regions contained very few tweets, which limit the kind of insight one could gain from this metadata. Therefore, we leveraged the domain expertise of former military intelligence officers at IDA to help us understand the interrelation between language and region. These domain experts have both foreign language expertise (particularly Russian) and geopolitical insight (from their years spent living and working in relevant countries):

- For those supposed regions from which few tweets were posted (e.g., Czech Republic, Denmark, and so forth), the IDA domain experts binned the regions into groups, based on their hypotheses of the IRA's geopolitical intent towards these regions (i.e., why the IRA might be tweeting from a certain region in a particular language). For example, the Czech Republic, Denmark, Egypt, Finland, Greece, and Latvia were all binned together, due to the similarities in their strategic and diplomatic relationships to the Russian Federation.
- In contrast, France and Hong Kong were not included in that bin, due to their unique relationships with the Russian Federation.

Although these bins were not used in this particular analysis, they could prove useful for future studies.

REPORT DOCUMENTATION PAGE				Form Approved OMB No. 0704-0188	
The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.					
1. REPORT DATE January 2020		2. REPORT TYPE FINAL		3. DATES COVERED (From-To)	
4. TITLE AND SUBTITLE Unsupervised Machine Learning Reveals How the Twitter Topic Patterns of the Russian-Backed Internet Research Agency (IRA) Evolved Over Time (Presentation)				5a. CONTRACT NUMBER HQ0034-14-D-0001	
				5b. GRANT NUMBER	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S) Parrish, Emily M. Cazares, Shelley M. Holzer, Jenny R.				5d. PROJECT NUMBER CRP 2234	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Institute for Defense Analyses 4850 Mark Center Drive Alexandria, VA 22311-1882				8. PERFORMING ORGANIZATION REPORT NUMBER IDA Document NS D-12056	
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) Institute for Defense Analyses 4850 Mark Center Drive Alexandria, VA 22311-1882				10. SPONSOR/MONITOR'S ACRONYM(S) IDA	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release; distribution is unlimited (10 March 2020).					
13. SUPPLEMENTARY NOTES					
14. ABSTRACT BACKGROUND: Information Operations (IO) are a key component of our adversaries' strategy to undermine U.S. military power without escalating to more traditional (and more easily identifiable) military strikes. Social media activity is one method of IO. In 2017 and 2018, Twitter suspended thousands of accounts likely belonging to the Kremlin-backed Internet Research Agency (IRA). Clemson University archived a large subset of these tweets (2.9M tweets posted by over 2800 IRA accounts), tagged each tweet with metadata (date, time, language, supposed geographical region, number of followers, etc.), and published this dataset on the polling aggregation website FiveThirtyEight. METHODS: Machine Learning researchers at the Institute for Defense Analyses (IDA) downloaded Clemson's dataset from FiveThirtyEight and analyzed both the content of the IRA tweets and their accompanying metadata. Using unsupervised learning techniques (Latent Dirichlet Allocation), IDA researchers mapped out how the patterns in the IRA's tweet topics evolved over time. RESULTS: Results showed that the IRA started tweeting in/before February 2012, but ramped up significantly in May/June 2015. Most tweets were in English, and most likely targeted the U.S. The IRA created new accounts after the first Twitter suspension in November 2017, with each new account quickly establishing an audience. Between at least January 2015 and October 2017, the IRA's English tweet topics evolved over time, becoming tighter, more specific, more negative, and more polarizing, with the final pattern emerging in late 2015. DISCUSSION: The United States government must expect that our adversaries' social media activity will continue to evolve over time. Efficient processing pipelines are needed for semi-automated analyses of time-evolving social media activity.					
15. SUBJECT TERMS Information Operations (IO); Internet Research Agency; Latent Dirichlet Allocation (LDA); Russian; social media; Topic Model; Twitter; Unsupervised Machine Learning					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT SAR	18. NUMBER OF PAGES 131	19a. NAME OF RESPONSIBLE PERSON Buckley, Leonard J.
a. REPORT Uncl.	b. ABSTRACT Uncl.	c. THIS PAGE Uncl.			19b. TELEPHONE NUMBER (include area code) (703) 578-2800