INSTITUTE FOR DEFENSE ANALYSES

Weaponized Tweets

(Presentation)

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

June 2021

IDA Document NS D-22805

Log: H 21-000333

INSTITUTE FOR DEFENSE ANALYSES
730 East Glebe Road
Alexandria, Virginia 22305-3086

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, project C2234 "Machine Learning for Social Media." 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

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

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

Executive Summary

Beginning in Summer 2018, the Institute for Defense Analyses (IDA) devoted limited internal research and development funds to exploring foreign malign influence operations in social media. To date, we have focused on the Twitter activity of the Russian-backed Internet Research Agency (IRA) in the run-up to the 2016 American Presidential election. We used a data-driven, quantitative approach with human-in-the-loop, employing unsupervised Machine Learning techniques (Latent Dirichlet Allocation) and other data analytic methods. We focused on the evolution of the adversary activity, tracking how topic patterns in that activity changed over time from 2012 through 2018. Our results uncovered multiple distinct tactical phases of the IRA attack. The IRA first began targeting Americans in late 2014 to early 2015. Their English tweet topics grew more specific over time, more varied, more negative, and more polarizing. Their final tweet topic pattern first emerged around late 2015, one full year before the 2016 election.

Weaponized Tweets

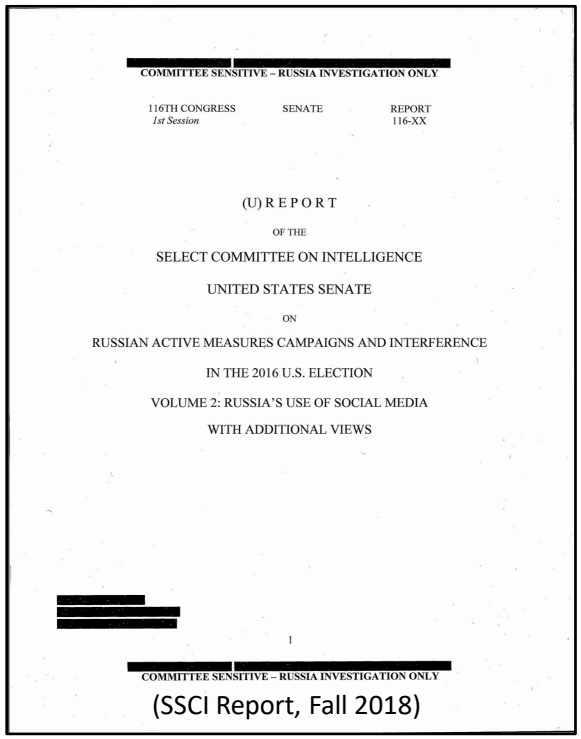
Shelley Cazares, Emily Parrish, Jenny Holzer
Science and Technology Division
Institute for Defense Analyses

June 2021

The Institute for 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.

Beginning in Summer 2018, IDA devoted limited internal research and development funds to exploring foreign malign influence operations in social media. Specifically, to date, we have focused on the Twitter activity of the Russian-backed Internet Research Agency (IRA) in the run-up to the 2016 American Presidential election. We used a data-driven, quantitative approach, employing unsupervised machine learning techniques and other data analytic methods. We focused on the *evolution* of the adversary activity, tracking how topic patterns in that activity *changed over time* from 2012 through 2018. These slides summarize our analysis.

IDA Monday Morning Quarterbacking?



[redacted] COMMITTEE SENSITIVE – RUSSIA INVESTIGATION ONLY

116TH CONGRESS *1st Session* SENATE REPORT 116-XX

(U) 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

[redacted]

1

[redacted] COMMITTEE SENSITIVE – RUSSIA INVESTIGATION ONLY

(SSCI Report, Fall 2018)

"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 2116."

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

Contact: scazares@ida.org

It is now common knowledge, in 2021, that the Russian-backed IRA has long sought to meddle online in American culture and politics, using social media as the new battleground for influence operations:

- The bipartisan Senate Select Committee on Intelligence (SSCI) published a comprehensive report in Fall 2018 (SSCI 2018).
- The former acting director of the CIA lamented about it in Summer 2019, saying, "Had we identified it much earlier, say in 2012–2013, we could have had more options… in the Summer of 2116" (Morell and Bharara 2019).

That is easy to say now, with perfect hindsight.

But, is that just Monday morning quarterbacking? How easy would it have actually been, prior to the 2016 election, to generate indications and warnings that the Russians were launching this kind of attack?

That is the question that we set out to answer at IDA, starting back in 2018. We wanted to put ourselves into the mind of a counter-intel analyst, starting back in the early 2012 time frame. To that end, we set up the following analysis as a thought experiment: Hypothetically speaking, imagine it is 2012 and a U.S. counter-intel unit gets a high-confidence tip that an adversary nation state is clandestinely posting tweets on social media. And let's say that source even identifies the particular Twitter accounts—those that, years later, Twitter itself will eventually attribute to the Russian-backed IRA. But back in 2012, imagine that a counter-intel analyst is tasked to monitor those flagged accounts for indications and warnings that the adversary *might be* using social media in some *nefarious* way against the United States:

- What does that analyst *do*?
- What open source information would that analyst have had easily available at the time?
- What technical tools would she have had easily available?
- And how could she have *used* those tools to generate and deliver timely, high-confidence indications and warnings to decision makers, in near-real time?

That is what this presentation will talk about today.

The story we tell is about a *domestic* problem—foreign adversaries meddling in U.S. culture and politics. However, this same type of approach could be used to address *foreign* problems—adversaries meddling in our allies' culture and politics, straining our relationships with those allies, and threatening U.S. military strength and national security.

IDA Bottom Line Up Front

- We created a prototype system using **open source software tools** to help analysts track the activity over time of known adversaries on social media
- Our system was not intended to replace the analyst, but rather help them perform their tradecraft at scale, providing evidence that:
  - The **Russian-backed Internet Research Agency (IRA)** began targeting English-speaking Americans in late 2014/early 2015
  - Throughout 2015, the IRA's English tweet topics **grew more specific, more varied, and more polarizing over time**
  - The IRA's final pattern **emerged in late 2015**, one year before the 2016 election
- Moving forward, we envision a **dashboard system for decision support**, allowing for a human-centered, multi-disciplinary, drill-down approach to allow intel analysts to adapt along with our adversaries

Contact: scazares@ida.org

A quick spoiler alert:

- We created a prototype system using open source software tools that could help intelligence analysts track adversary activity in social media—in particular, to track how the adversary's activity *evolves over time*. Then we used our tools to analyze the Russian-backed IRA's Twitter activity, starting in 2012 and continuing up into 2018.
- Our system was not intended to *replace* the analysts, but rather *aid* the analysts in performing the same type of tradecraft they always do, but applying it to millions of tweets, rather than just a few dozen documents.
- Our results show that the IRA's English tweet topics did *not* remain constant in the run-up to the 2016 election. Instead, they *evolved over time*. In fact, we uncovered *multiple distinct tactical phases* of their attack, all of which could have been provided to U.S. government decision makers with only one month of lag time. To be specific, the IRA first began targeting Americans in late 2014 to early 2015. Their English tweet topics grew more specific over time, more varied, more negative, and more polarizing. Their final pattern first emerged around late 2015, which happened to be one full year before the 2016 election.
- Moving forward, we envision a dashboard system for decision support, allowing for a human-centered, multi-disciplinary, drill-down approach that could allow analysts to adapt alongside our adversaries in real time.

IDA Methods for Real-Time Analysis

- **Twitter has attributed** thousands of accounts belonging to organizations involved with state-backed information operations, such as the IRA
- We downloaded a **publicly-available dataset** of ~3M tweets posted by ~3K IRA accounts between Feb 2012 – May 2018
- We simulated a near-real-time analysis:
  - We cleaned and **binned the tweets by month**
  - We **fit a topic model** to each month's tweets
  - The topic model organized the words of each month's tweets into underlying **topics**
  - An analyst could have quickly glanced through these topics to get a sense of what issues the IRA was promoting on Twitter

This semi-automated approach could help intel analysts scale their tradecraft to the huge numbers of posts on social media.

Contact: scazares@ida.org

First, we will provide an overview of our methods:

- Over the past few years, social media platforms like Twitter have been working with U.S. authorities to identify active measures campaigns. Twitter then packages up and releases the posts made by these adversarial accounts, so that researchers like us can study and learn from them. (Twitter Transparency Report 2020).
- We began our analysis at IDA in Summer 2018, when Twitter first released the account names (handles) of over 3000 accounts belonging to the IRA. Clemson University packaged up about three million of the tweets posted by these flagged accounts between early 2012 to mid 2018, and then published them on the data aggregation website *FiveThirtyEight* (Roeder 2018). We downloaded them the very next day.
- Our technical approach simulated a near-real-time analysis—what *could* have been done, starting back in the 2012 timeframe:
  - First, we cleaned the tweets to remove non-essential characters like punctuation, emojis, and linked URLs. Then we binned the tweets by month.
  - Next, we fit a statistical *topic model* to each month's worth of tweets.
  - The topic model organized the words of each month's tweets into underlying *topics*. From a technical perspective, topics are lists of words that often occur together in the same tweets.
  - In real time, an analyst could have quickly glanced through these topics to get a sense of what issues the IRA was promoting on Twitter in any given month.

This semi-automated approach could help intel analysts *scale their tradecraft* to the huge numbers of posts on social media.

IDA Topic Modeling with Latent Dirichlet Allocation (LDA)

- 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`

Contact: scazares@ida.org

To perform our topic modeling, we used a well-known approach called Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003). LDA fits a statistical model to a corpus of documents, clustering the topics. In our analysis:

- A *corpus* is the term we use to describe 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 our term for 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 statistical mixture of the topics.
- A *topic* is a list of words that are found to be statistically associated with each other because they frequently occur together in the same tweets (e.g., isis syria targeted targets Iraq forces accounts Israel u.s mosul opiceisis iceisis aleppo north killed yemen refugees airport Syrian saa).

LDA was developed in the early 2000s and has been implemented in a variety of freely available software packages and code libraries. We used the MALLET software (McCallum 2002) to performance LDA. We chose to use this software fresh out the box, with very little deviation from its default parameter values. In that sense, our findings outline the *lower* bound of capability—what kind of patterns or actionable intel one could *easily* extract from tweets like these, even back in the 2012–2015 time frame.

We then summarized each topic. In our analysis:

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

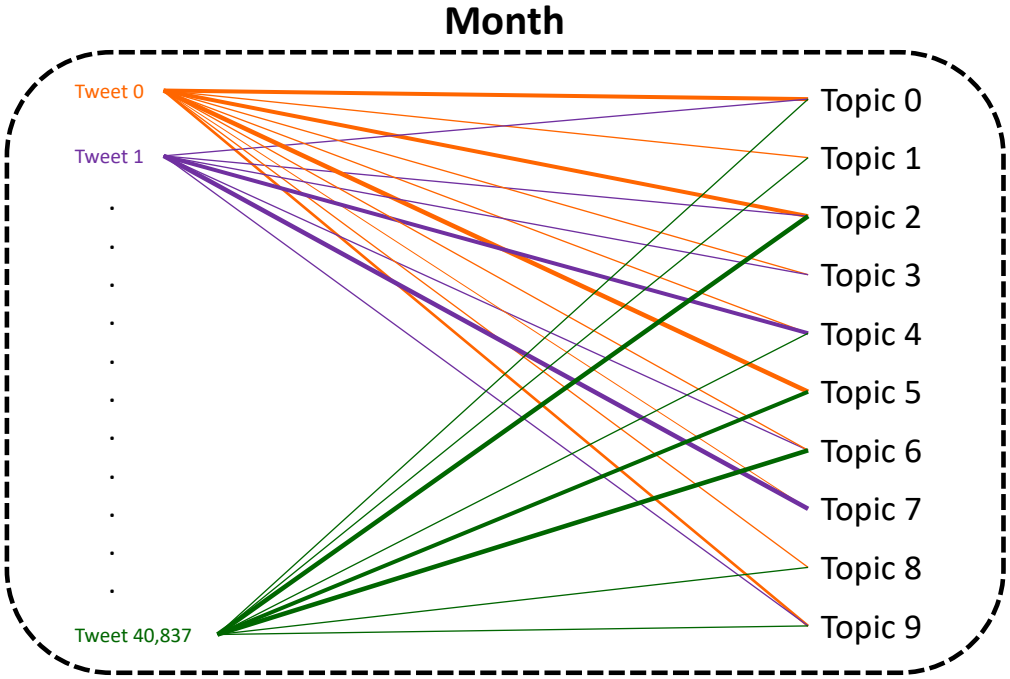
Human intelligence is needed to manually perform this final, topic labeling step. To accurately interpret and summarize the long lists of words for each topic, the human labeler requires expertise in (1) the *language* in which the tweets were written (in this case, English) and (2) the *social and cultural themes* that Twitter users were likely to have encountered (in this case, in the United States in the mid 2010s). We served as our own human topic labelers for the purposes of this analysis.

(Note that Twitter allows other users to comment on any given tweet. We did not include the comments to a tweet as part of the tweet's content. Twitter considers each comment to be a tweet in and of itself. For each tweet in our analysis:

- Those comments that were posted by "regular people", i.e., *non*-known-IRA accounts, were *not* considered in our analysis.
- Those comments that were posted by known IRA accounts were treated as separate tweets in our analysis.

Unfortunately, the dataset we used did not explicitly keep track of which IRA tweets were comments to other IRA tweets, and so we could not track how a particular message propagated through Twitter, on a comment-by-comment basis. However, the dataset did keep track of how many followers an IRA account had when they posted a tweet—which we discuss on a later slide.)

IDA Topic Modeling with Latent Dirichlet Allocation (LDA)



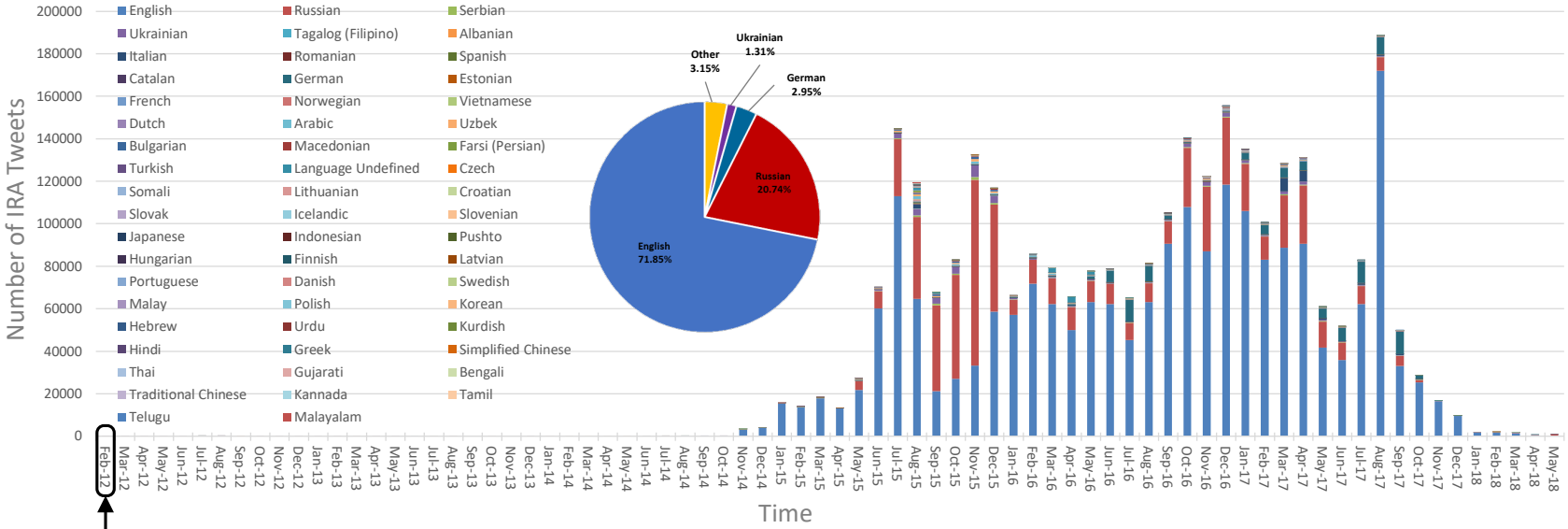
LDA allows each document (tweet) to have multiple topics.

Contact: scazares@ida.org

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. We set up the LDA statistical model to assume 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 statistical association to each topic, represented here as the thickness of each tweet-to-topic line. For example, the final Tweet #40,837 (green) is most strongly associated with Topic #6, followed by Topic #2. This final tweet is hardly associated at all with Topics #3 or #7.

The mathematical equations underpinning the LDA statistical model are described in the backup slides.

IDA **Feb 2012:** IRA Tweets over Time, by Language



Few IRA tweets, with even fewer in English.

Feb 2012

Contact: scazares@ida.org

Now we will summarize our results.

We plot the number of IRA tweets in the dataset versus time, which is binned by month. The color denotes the language that the tweet was posted in.

Automated software tools could have identified the most prominent languages of these tweets, to display simple pie charts like this. The pie chart shows that, overall from February 2012 through May 2018, most of the tweets were posted in English (71.85% overall, medium blue), presumably indicating that the IRA mostly targeted an English-speaking audience. Other prominent languages were Russian (20.74% overall, red), German (2.95% overall, dark blue), and Ukrainian (1.31% overall, purple). All other languages collectively comprised 3.15% of the total (yellow).

However, it's important to look at the *evolution* of the tweet languages over time. The earliest tweets in the dataset were posted in February 2012. In the first two years of the dataset, few IRA accounts were active, each posting only several dozen to a few hundred tweets each month—so few that it is difficult to see their bars on this vertical scale. Furthermore, most tweets were in Russian and only a few were in English—although it's difficult to see at this scale, most of the little short bars on the left edge of the plot are red, indicating most of the early tweets were posted in Russian.

IDA **Feb 2012:** An Early IRA Tweet

@iris0_O
Реклама в Facebook: причина успеха: Facebook сегодня, как ожидается, подаст заявку на IPO. Размещение ожидается ... http://t.co/Ovdc6zNv
12:35 AM • Feb 2, 2012 from United States

Translation

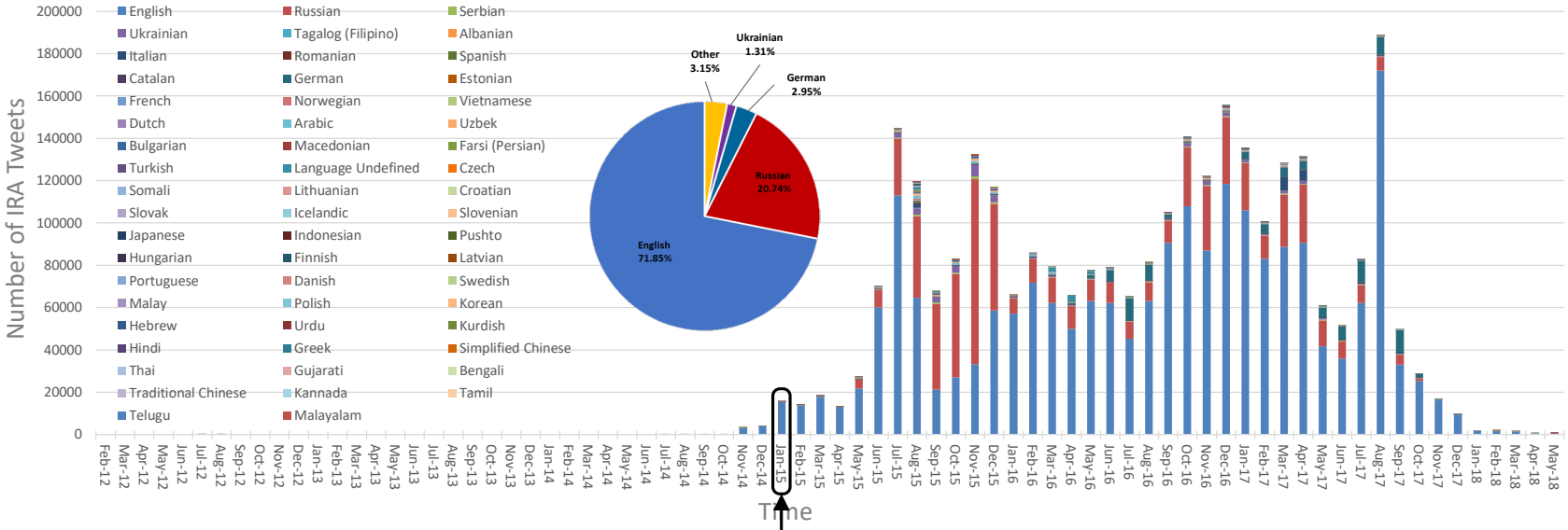
@iris0_O
Advertisement on Facebook: Reason for success: Facebook today, as expected, will announce their IPO. The announcement is expected ... http://t.co/Ovdc6zNv
12:35 AM • Feb 2, 2012 from United States

Few IRA tweets, with even fewer in English.
Analyst could have read each tweet manually.

Contact: scazares@ida.org

With this manageable, fairly small number of tweets, a counter-intel analyst could have read each one manually. Glancing through a sea of Cyrillic characters, like this tweet here, the analyst would would have quickly surmised that most tweets weren't meant for an English-speaking American audience. She could have passed off these tweets to the appropriate language or regional experts and settled into a sit-and-watch mode, for the next two years.

IDA **Jan 2015:** IRA Tweets over Time, by Language



Significantly more IRA tweets, most in English.

Jan 2015

Contact: scazares@ida.org

Things changed in late 2014 and early 2015. Simple counts show that the IRA began tweeting much more frequently at this point in time, by over two orders of magnitude—the bars of this chart start getting easier to see at this vertical scale. By January 2015, English (medium blue) had become the most prominent language of the IRA. An analyst would have sat up and noticed:

Why are the flagged accounts posting so frequently and so heavily in English?

Are the Russians now targeting Americans?

The analyst would not have been able to read through the tens of thousands of flagged tweets. Instead, she could have let her computer do the reading for her, with LDA's topic modeling supporting her workflow by automatically organizing the topics of the tweets for her to label, thus providing a semi-automated change detection on the Russians' observable tactics.

IDA **Jan 2015:** Three IRA English Topics & Example Tweets

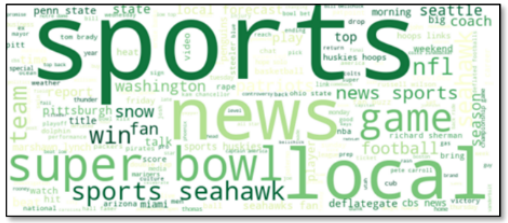
Motivation

@stanley_cox_
What's life worth living with frown? Pick your head up and put a smile on your face and never forget what makes you happy.
2:07 PM • Jan 1, 2015 from United States

Local News

@ScreamyMonkey
#News #Crime Contra Costa County foster youth's mysterious death under investigation (San Jose Mercury News) http://t.co/TAlQH3vsK2
14:32 PM • Jan 13, 2015 from United States

Sports



@ScreamyMonkey
#News #US Patriots coach Bill Belichick insists team followed rules in "deflate-gate" (CBS News) http://t.co/ai162QbE9E
10:35 PM • Jan 24, 2015 from United States

Significantly more IRA tweets, most in English.
Topics were loose, vague, and both positive and negative.

Contact: scazares@ida.org

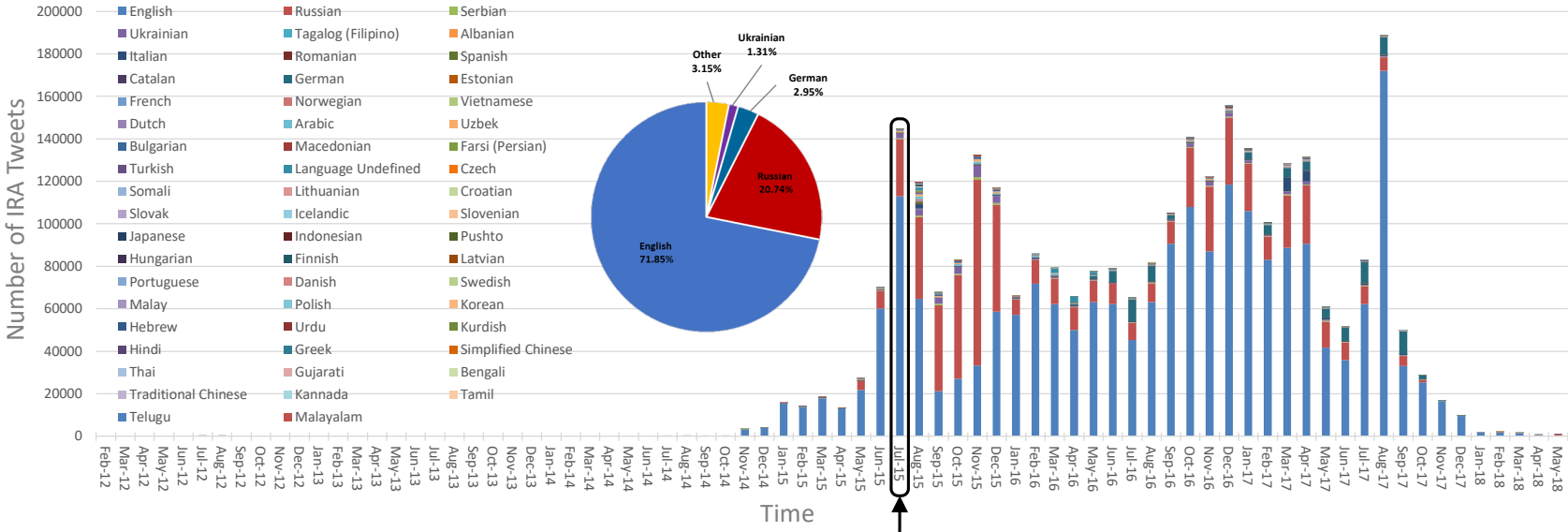
As mentioned earlier, we used LDA to fit 10 topics to our data for each month. Here we show *three* of the strongest topics for January 2015—three topics that were heavily represented in the statistical model fit to the January 2015 tweets. We also show a representative tweet for each topic—that tweet for which the probability of that topic was highest or very high:

- Left: Many of the tweets in this month used words like *make*, *love*, *time*, *thing*, *success, give*, and *smile*—loose, vague, and positive terms that an analyst could have labeled "Motivation."
- Center: Also, many tweets in this month used words like *news*, *local*, *police*, *fire*, *killed*, and so on—similarly loose and vague but negative terms that could have been labeled "Local News."
- Right: And many tweets in this month used words like *sports*, *NFL*, *super*, *game*, and *bowl*—a "Sports" topic.

Based on this information, the analyst could have reported that the *IRA's tactics had changed*: They were suddenly posting much more frequently, in English, about loose, vague topics with both positive and negative affect—motivational messages, local news, and sports. By objectively aggregating a very large number of tweets, the system could have allowed the analyst to provide this high-confidence assessment of the IRA's observed tactics.

However, inferring the IRA's *strategy* based on those observed tactics would have still been a more difficult task in real time.

IDA **Jul 2015:** IRA Tweets over Time, by Language



Even more IRA tweets, most still in English.

Jul 2015

Contact: scazares@ida.org

The situation changed again a few months later. In July 2015, the number of IRA tweets shot up by another order of magnitude. The percentage in English (medium blue) remained high.

IDA **Jul 2015:** Three IRA English Topics & Example Tweets

Local News

@TodayCleveland
Missing Hudson man found dead; man dies after Shaker Boulevard crash; LeBron James movie premieres: Northeast Ohio News Links ... #crime
6:16 PM • Jul 10, 2015 from United States

Politics

@TodayCleveland
Read Ohio Gov. John Kasich's announcement that he's running for president: Ohio Gov. John Kasich ann...
http://t.co/E0PpzA9V9Q #politics
9:03 PM • Jul 21, 2015 from United States

Exercise

@ExQuote
http://t.co/nh6MK3IJet GREAT class abs workout getfit move!
@[redacted] @[redacted]
@[redacted] @[redacted]
@[redacted] @[redacted]
9:42 PM • Jul 12, 2015 from United States

Even more IRA tweets, most still in English.
Topics dominated by a single cultural topic (exercise), interspersed with news & politics.

Contact: scazares@ida.org

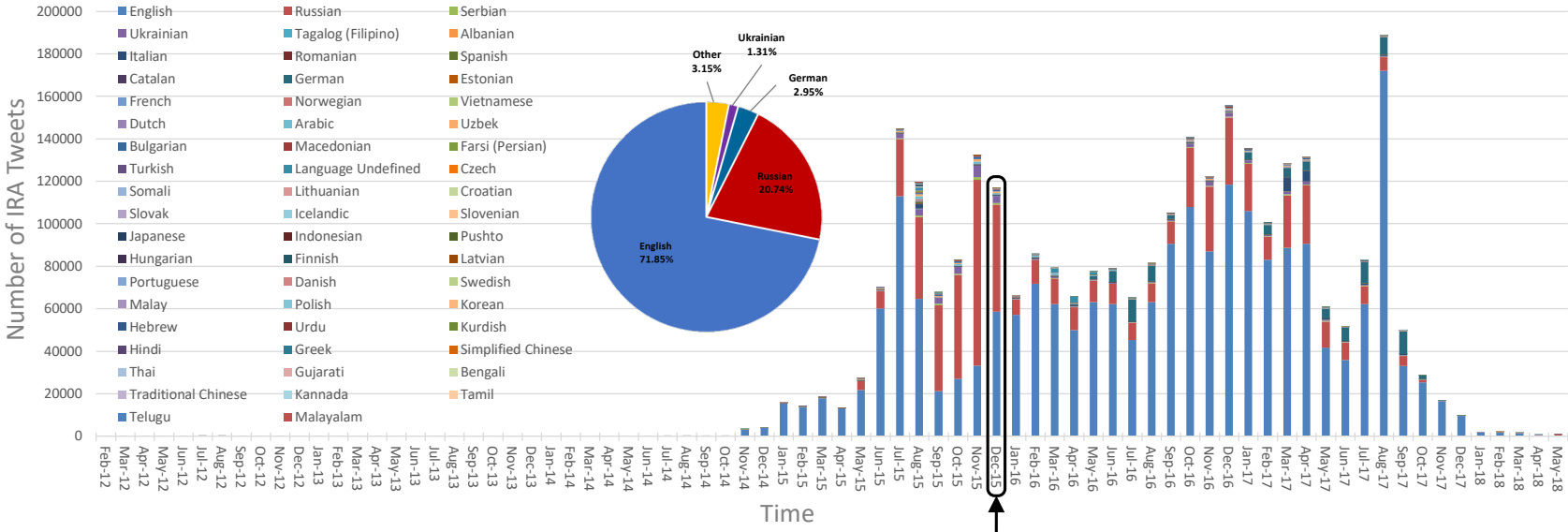
Once again, we show three of the strongest topics for this month, July 2015:

- Left: Some English tweets were associated with the same "Local News" topic from January 2015.
- Center: Another topic consisted of a hodgepodge of words, some political in nature.
- Right: And a large number of tweets were associated with a new topic: "Exercise." It's possible that the IRA may have used these "Exercise" tweets to pique interest from other Twitter users and garner a following (Weisburd, Watts, and Berger 2016). In fact, many "Exercise" tweets tagged the handles of other Twitter accounts, blacked out for privacy here.

At this point in time, July 2015, the analyst could have reported that the IRA had changed their tactics again: from loose, vague topics in *early* 2015 now to a single cultural topic (exercise) interspersed with news and some politics in the *summer* of 2115.

However, the *strategy* behind this latest set of tactics may have still remained unclear.

IDA **Dec 2015:** IRA Tweets over Time, by Language

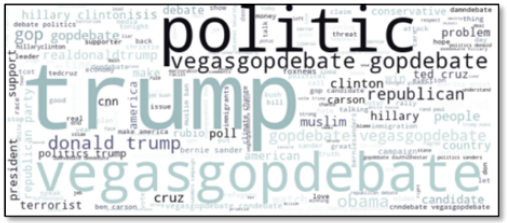


Contact: scazares@ida.org

Fast-forward to the *end* of 2115. The number of English tweets remained fairly high.

IDA Dec 2015: Three IRA English Topics & Example Tweets

Politics



@ClydeDawson

Limbaugh Warns Of 'Heart Attacks & Slit Wrists' In GOP Establishment & Media If Cruz Wins Iowa [AUDIO] https://t.co/D6AaSpH14z #CCOT #IOWA

4:09 PM • Dec 8, 2015 from United States

Guns

@TraceyHappyMom

The Republican Fear of Facts is the patriot faith flag Obama Hillary Benghazi gun with the Acorn Solyndra Obama gun Obama gun Obama gun gun

2:38 PM • Dec 24, 2015 from United States

Terrorism

@SouthLoneStar

U.S. visa process missed San Bernardino wife's online zealotry. That's the outcome of Obama's policies https://t.co/nSNJs3aDnh #Trump2016

11:17 PM • Dec 18, 2015 from United States

Most IRA tweets still in English.
Topics are more specific, more varied, more negative, and more polarizing.

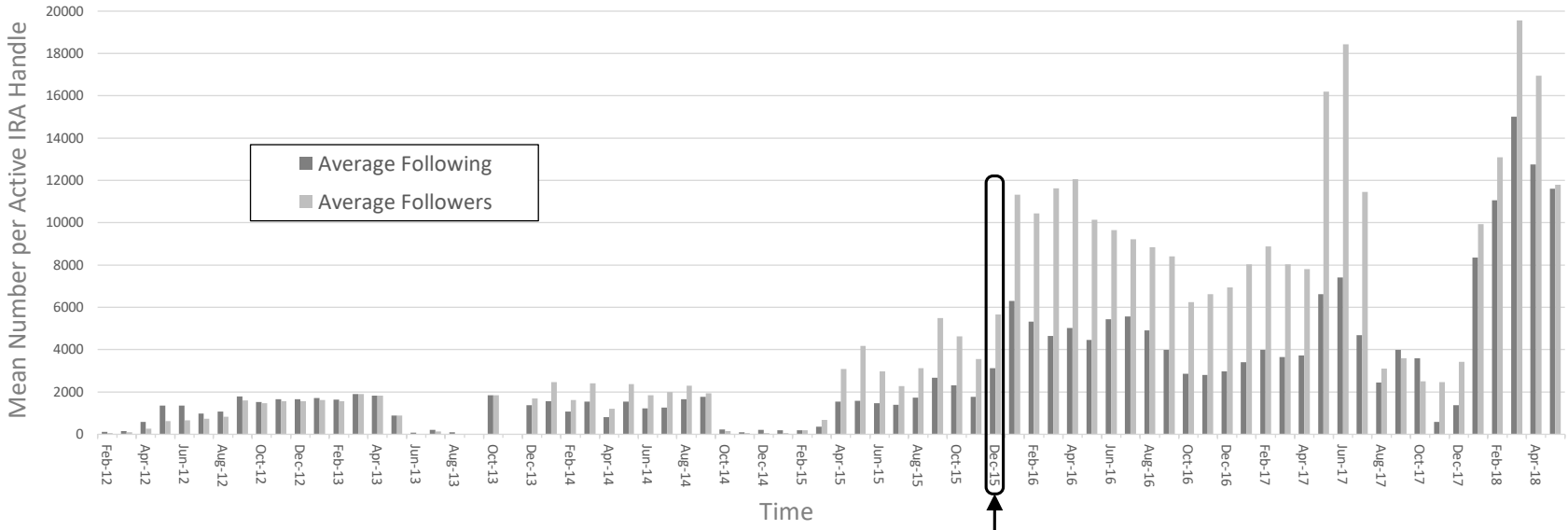
Contact: scazares@ida.org

Now, in December 2015, the IRA's topic pattern changed again—topics were now more specific, more varied, more negative, and more polarizing. Many tweets were still associated with the "Local News" and "Politics" topics. By now, though, like shown on the left, many "Politics" tweets had adopted a more polarizing tone. Other topics also began to emerge, including "Guns" and "Terrorism," on the center and right.

The fall of 2115 (one year before the 2016 election) was the earliest point in time in which our system's indications and warnings could have helped reveal Russia's strategy to sow discord in the English-speaking world (Senate Select Commit on Intelligence 2018).

This result would not have been a surprise—the Russians have wielded active measures against the United States and our allies since the Cold War (Weisburd et al. 2016). (For example, during the 1980s, the Soviets launched a disinformation campaign to spread the (false) narrative that AIDS was the result of secret U.S. military experiments (Kramer and Selvage 2020).)

IDA **Dec 2015:** Followers & Followings over Time



Followers increased significantly after late 2015.

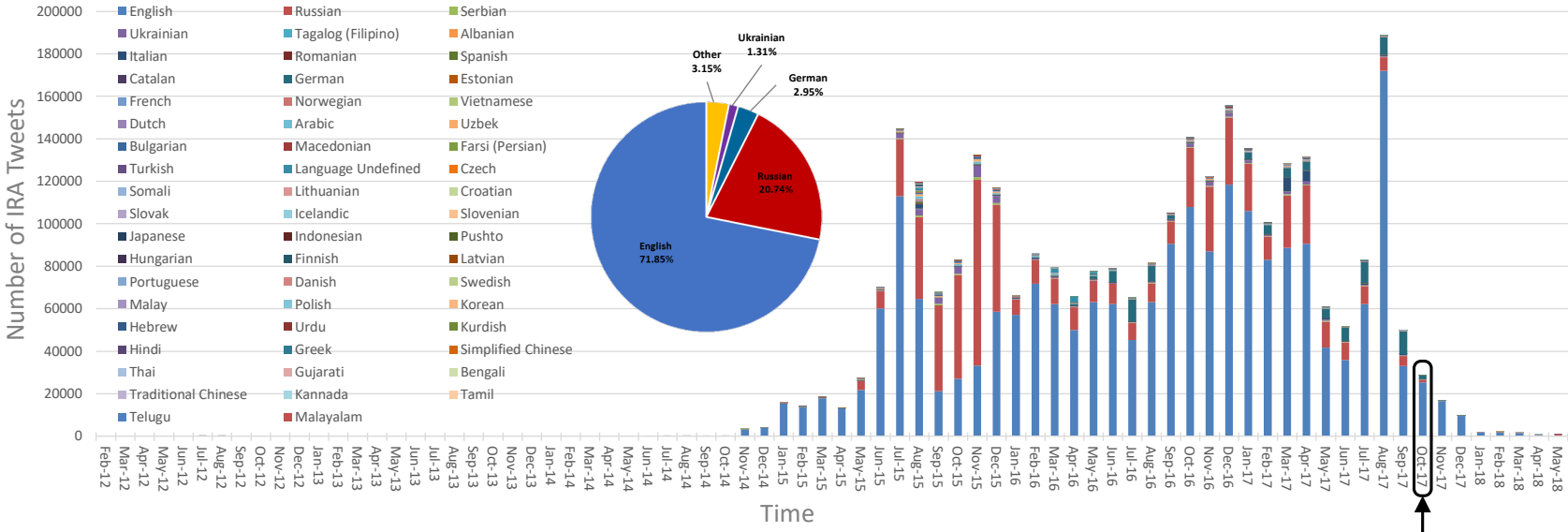
Dec 2015

Contact: scazares@ida.org

And interestingly enough, we see that this new set of tactics—this new set of topic patterns—also occurred shortly before a large increase in *followers*. On this slide, we plot the average number of following and follower accounts per active IRA handle over time. On average, the 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).

Looking at the light grey bars, we see a large increase in followers in early 2016, just a few months after the IRA settled into what has now become known as their tried and true pattern of topics. It may be that the IRA learned that specific, varied, negative, and polarizing topics led to more followers. Of course, correlation does not mean causation—therefore this is a *hypothesis* that needs to be explored in more detail, in order to understand how our adversaries measure their own impact and how they become incentivized to change their own tactics.

IDA **Oct 2017:** IRA Tweets over Time, by Language



Fewer IRA tweets, most still in English.

Oct 2017

Contact: scazares@ida.org

Regardless, the Russians' overall tactics remained largely unchanged after that, in terms of their topic patterns. In October 2017 (one year after the election), their tweet numbers had decreased, but the percentage in English was even higher than before.

(Of particular note is the large spike in tweets in August 2017. This was a particularly active month in current events, including the white nationalist protest in Charlottesville, VA (Stolberg and Rosenthal 2017) and the devastation of Hurricane Harvey in Houston, TX (National Weather Service n.d.). Further analysis is needed to better understand this spike in tweets.)

IDA **Oct 2017:** Three IRA Topics & Example Tweets

National Anthem Protests

@BarbaraForTrump
Trump is scorned by degenerate rappers, distorted Hollywood, and cop-hating NFL players. Makes me proud that I voted for Trump!
#trump45
11:45 PM • Oct 15, 2017

Hurricane Maria/Puerto Rico

@evannghartr
BOMBSHELL CLAIM: San Juan Mayor SABOTAGING Relief Efforts to Make Trump Look Bad
https://t.co/t9BROQ9zpj
https://t.co/DfMERR82OE
10:31 AM • Oct 1, 2017

Vegas Shooting

@evannghartr
BOMBSHELL VIDEO Raises Doubts on Official Vegas Story, Shows SECOND SHOOTER?
https://t.co/P0jiPj6Noi
https://t.co/heqiB7wr40
10:30 PM • Oct 3, 2017

Fewer IRA tweets, most still in English.
Topics are even more specific, varied, negative, and polarizing.

Contact: scazares@ida.org

In October 2017, the IRA's English tweet topics were still specific and varied, and now even more negative and polarizing. Topic labels now included "National Anthem Protests," "Hurricane Maria/Puerto Rico," and "Vegas Shooting":

IRA Twitter Topics Evolved over Time

Legend: Tweet Topic Category

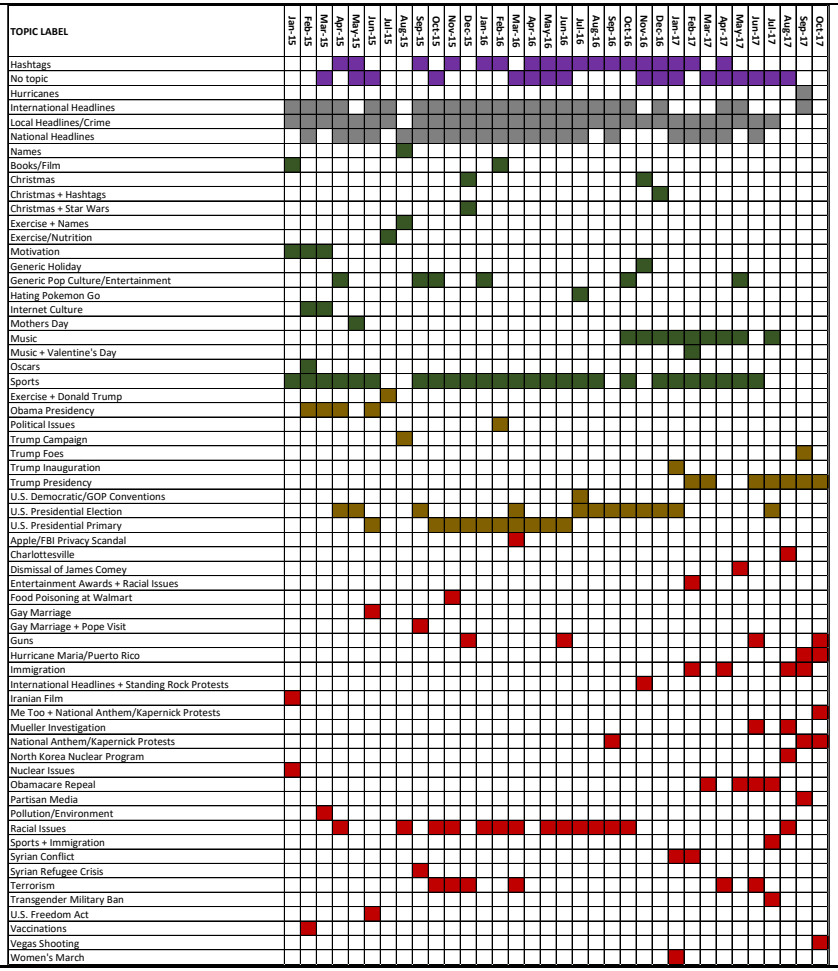
- Other
- News
- Entertainment/Culture
- U.S. Politics
- Polarizing Issue

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)



Contact: scazares@ida.org

This matrix summarizes *all* topics that appeared in the IRA English tweets over our period of topic analysis, from January 2015 through October 2017. Each column corresponds to a month, while each row corresponds to a topic. A square is filled with a color if that topic was represented in that month and is left blank otherwise. It may be difficult to read the individual topics, but 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.

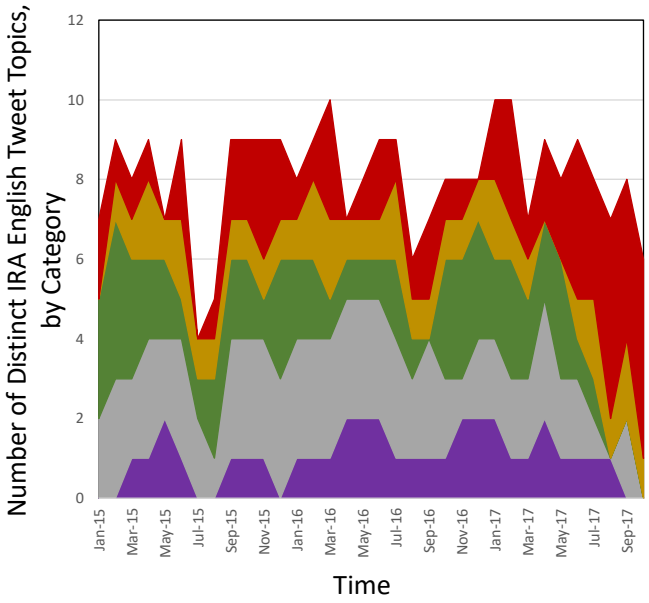
A quick glance shows that many of the IRA's English tweet topics were short-lived. Naturally, many of these short-lived topics referred to events with limited time spans, such as the Trump Inauguration. Furthermore, other short-lived topics referred to brief fads, such as Pokemon Go. However, other topics 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 News Headlines, Music, Sports, the U.S. Presidential Election/Primary, and Racial Issues. We hypothesize that the IRA used the News Headlines, Music, and Sports topics to maintain cover and attract a large American followership, while the Racial Issues topic was used to sow discord among its American followers, which was then directed towards political events and issues such as the U.S. Presidential Election/Primary.

IRA Twitter Topics Evolved over Time

Legend: Tweet Topic Category
- Other
- News
- Entertainment/Culture
- U.S. Politics
- Polarizing Issue

Most early topics were green, while most later topics were red.



Contact: scazares@ida.org

We can condense the matrix from the previous page into a simplified graph on this page. This graph plots the number of distinct topics that the IRA tweeted in each month. The color indicates the tweet topic category (Other, News, Entertainment/Culture, U.S. Politics, and Polarizing Issue). For example, the left-most edge of this graph shows that in January 2015, there were seven distinct topics. (Although LDA automatically organized that month's tweets into 10 topics, the human analyst found that three of those topics shared the same labels with others.) Two of the seven distinct topics related to News (grey), three more related to Entertainment/Culture (green), and the final two related to a Polarizing Issus (red). No topics in January 2015 were related to U.S. Politics (yellow) or fell into the Other category (purple).

A quick glance at the color shadings provides insight on how the IRA's English tweet topics evolved over time. In early 2015, towards the left of the matrix, most topics were shaded in green for Entertainment/Culture. Towards the end, on the right, most topics were red for Polarizing Issues. The IRA's English tweet topics grew more polarizing, with the first rise of the polarizing issue topics occurring in late 2015, one full year before the 2016 election.

IDA Conclusions

- Our adversaries can and have adapted over time. We must, too.
- Machine learning systems must scale.
- Social media influence operations can occur at any time.
- Social media influence operations are a multi-disciplinary issue, involving:
  - Science & Technology
  - Geopolitics & Foreign Language
  - Social Science

Contact: scazares@ida.org

Our analysis led to four main conclusions:

- First, just because our adversary's tactics (i.e., topic patterns) *haven't* changed much in the last couple of years doesn't mean they *won't* change a lot in the future. They *did* change a lot in just 2015 alone—from loose and vague Motivation and News topics in *early* 2015, to single-topic Exercise tweets in *mid* 2015, to specific and varied and negative and polarizing topics in *late* 2015. We know our adversaries *can* adapt because they *have* adapted. And so, *the United States* must be able to adapt, too. To address this time evolution requirement, the U.S. government needs machine learning systems that can adapt over time, if *and when* our adversaries adapt.
- Second, any machine learning system for social media influence operations has to scale to a huge number of posts. For the months with the most tweets, our system handled tens of thousands of tweets and could have easily handled more. It was the human that was the bottleneck. Our human topic labeler had to read through the topic word lists for each month and apply semantic labels to them. That was tricky and took some time. Since then, we've worked out a Standard Operating Procedure and created more high-tech visualization tools to make that easier, faster, and more systematic for a human analyst. However, it still takes some time and discipline. Furthermore, if we wanted to do the topic labeling on a week by week basis, or even a day by day basis, could a human analyst keep up with that pace? Probably not with right-out-of-the-box LDA techniques. Other machine learning methods (Blei and Lafferty 2006) could potentially automate a bit more of the topic labeling process, lightening the load on the human analyst. (That is, other methods could potentially relax some of the independence assumptions in the underlying LDA equations to model the *statistical correlation between tweets over time*, automatically tracking which tweet topics continue from month to month.) Then maybe the analyst could respond more quickly to more tweets. (Further information about LDA and its independence assumptions is contained in the backup slides.)
- Third, these past couple of years have shown us that influence operations can happen at any time. Not just in the run up to elections, but for any event. Like a pandemic. Or a protest. A lot has been written about disinformation campaigns regarding the 2019 protests in Hong Kong or the 2020 coronavirus pandemic. Therefore, for the future, we envision a dashboard system for decision support. Anytime an analyst is tasked with monitoring accounts that have been attributed to foreign malign influence operation, the analyst could pose a series of queries to the system, to drill down into what topics our adversaries are promoting at specific points in time. (The backup slides contain our further thoughts about this kind of query system.)

- And finally, this is a multi-disciplinary problem, requiring expertise in at least three areas:
  - First, Science and Technology, to create the advanced systems that can automate more of the topic labeling process, leaving less load on the human analyst.
  - Second, Geopolitics and Foreign Language, to understand and provide context for the events that are happening in different regions of the world.
  - And third, Social Science. We can't forget the "social" in social media. Twitter is not just a bunch of bots tweeting at each other. There are real people involved who formulate tweets and interpret tweets and take action on tweets, all against the backdrop of their own cultural experiences.

Luckily the United States has all of these skill sets. The hard part is getting everyone together under one roof, and herding the cats.

IDA Further Information

War on the Rocks article

https://warontherocks.com/2020/10/weaponized-tweets-artificial-intelligence-could-help-defend-against-adversary-attacks-in-social-media/

IDA Ideas podcast

https://idaideas.podbean.com/e/ida-ideas-weaponized-tweets/

Discussion

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

Any Questions?

There is more information available about our work online. Please see our recent article in War on the Rocks (Cazares, Parrish, and Holzer 2020) and our podcast on the IDA website (Cazares, Parrish, Holzer, and Moeller 2020). Please also feel free to contact us directly over phone or email.

IDA References (1 of 2)

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

Blei, David M. and John D. Lafferty. (2006). *Dynamic Topic Models.* https://dl.acm.org/doi/pdf/10.1145/1143844.1143859

Blei, David. (2019). Topic Models. *VideoLectures.net*. http://videolectures.net/mlss09uk_blei_tm/

Cazares, Shelley, Emily Parrish, and Jenny Holzer. (2020). Weaponized Tweets: AI Could Help Defense Against Adversary Attacks on Social Media. *War On The Rocks.* https://warontherocks.com/2020/10/weaponized-tweets-artificial-intelligence-could-help-defend-against-adversary-attacks-in-social-media/

Cazares, Shelley, Emily Parrish, Jenny Holzer, and Rhett Moeller. (2020). Weaponized Tweets: Artificial Intelligence to Defend Against Influence Operations in Social Media. *IDA Ideas.* https://idaideas.podbean.com/e/ida-ideas-weaponized-tweets/

Kramer, Mark and Douglas Selvage. (2020). Lessons from Operation "Denver," the KGB's Massive AIDS Disinformation Campaign. *The MIT Press Reader.* https://thereader.mitpress.mit.edu/operation-denver-kgb-aids-disinformation-campaign/

McCallum, Andrew Kachites. (2002). *MALLET Machine Learning for LanguagE Toolkit.* http://mallet.cs.umass.edu/

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

National Weather Service. (No Date). *Hurricane Harvey & Its Impacts on Southeast Texas (August 25-29, 2017).* https://www.weather.gov/hgx/hurricaneharvey

Contact: scazares@ida.org

We list documents cited in these slides.

IDA References (2 of 2)

Rebello, Katarina, Christian Schwieter, Marcel Schliebs, Kate Joynes-Burgess, Mona Elswah, Jonathan Bright, and Philip N. Howard. (2020). *Covid-19 News and Information from State-Backed Outlets Targeting French, German, and Spanish-Speaking Social Media Users.* Oxford Internet Institute, University of Oxford. https://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2020/06/Covid-19-Misinfo-Targeting-French-German-and-Spanish-Social-Media-Users-Final.pdf

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

SSCI. (2018). *Report on 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.* https://www.intelligence.senate.gov/sites/default/files/documents/Report_Volume2.pdf

Stolberg, Sheryl Gay and Brian M. Rosenthal. (2017). Man Charged After White Nationalist Rally in Charlottesville Ends in Deadly Violence. *The New York Times.* https://www.nytimes.com/2017/08/12/us/charlottesville-protest-white-nationalist.html

Twitter Transparency Report. (2020). *Information Operations.* https://transparency.twitter.com/en/reports/information-operations.html

Weisburd, Andrew, Clint Watts, and J.M. Berger. (2016). Trolling for Trump: How Russia is Trying to Destroy Our Democracy. *War On The Rocks.* https://warontherocks.com/2016/11/trolling-for-trump-how-russia-is-trying-to-destroy-our-democracy/

Contact: scazares@ida.org

We list documents cited in these slides.

IDA LDA: A Probabilistic Generative Model

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 words in corpus to probabilistic model

Assume each document was drawn independently from that model

Assume each word in each document was drawn independently from that model

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

θ_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)

Contact: scazares@ida.org

This backup slide writes out the equation underpinning LDA (Blei et al. 2003, Blei 2019). LDA is a probabilistic generative model used for topic modeling of text-based documents. The term "generative" means that, once the model is fit to the data, it can be used to simulate how the data was generated in the first place.

LDA makes several assumptions about statistical independence:

- First, the blue part of the equation on the left explains that the words in a corpus $\mathcal{C}$ (i.e., a month) must be fit to a probabilistic model, given a hidden variable η, which is drawn from the Dirichlet distribution.
- Then, the green part of the equation in the center explains that we assume each document (i.e., each tweet) was drawn independently from the statistical model, given another hidden variable α, also drawn from the Dirichlet distribution.
- Finally, the rest of the equation on the right, mostly in yellow, explains that we also assume each word in each document (i.e., in each tweet) was drawn independently from the statistical model.

LDA differs from a standard mixture model:

- With standard mixture models, all words of a document (i.e., of a tweet) are assumed to have been drawn from the *same* topic. That is, each document (i.e., each 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., of each tweet) may have been generated by a different topic—or even from multiple topics. That is, each document (i.e., each tweet) may be associated with multiple topics.

More detail about this equation is included in the next slide.

LDA: A Probabilistic Generative Model

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

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

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

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

(1: Words in corpus; V; 1: Topics in corpus; 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):

$$\theta_d = \begin{bmatrix} & \cdots & \end{bmatrix}$$

(1: Topics in corpus; 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

θ_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)

Contact: scazares@ida.org

This backup slide provides more detail about the equation underpinning LDA (Blei et al. 2003, Blei 2019).

Like many probabilistic 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: wdn, zdn, θd, 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,jth element of β estimates the probability of the jth unique word in the ith 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).
- θd is a vector that keeps track of how strongly each document in the corpus is associated with each topic. Each element of θd refers to one of the K topics in the corpus. The ith element of θd estimates the probability of the ith topic in the dth document. θd is also a hidden variable; therefore, we must estimate it, as well. We do this given all words in the dth document (observables) and α (another hidden variable, also drawn from the Dirichlet distribution).
- zdn is the topic from which the nth word in the dth document is drawn. We can look up the probability of zdn from θd.
- wdn is the probability of the nth word in the dth document. We can use zdn to look up the probability of wdn from β.

Several different statistical techniques can be used to fit the elements of β and θd to the data, given α and η, respectively. We used the software package MALLET (McCallum 2002).

IDA 2020s: A Dashboard System for Decision Support

For all tweets posted by the foreign adversary accounts:

- Show me all that were posted:
  - In the last month,
  - In the German language,
  - From a supposed German location, and
  - Between 0200 and 0500 Central European Time (when most real Germans are asleep)…
- Of these, show me all that were related to the topic labels "Coronavirus Statistics" or "Social Distancing"…
- This tweet looks interesting. Show me all tweets like this one (with a similar combination of topics) from March 2020 (when Germany first imposed social distancing restrictions)…
- Of these, show me all tweets that included the hashtag #Wuhan…
- And so forth…

Contact: scazares@ida.org

This backup slides discusses our thoughts for a query-based system for social media analysis.

Machine learning applications like the topic-sorting software tool we discussed above could have helped analysts track changes in the Russians' observable tactics over time, thus providing early indications and warnings of their intent. Moving forward, how could the U.S. intelligence community rapidly and efficiently repeat this type of analysis in the 2020s and beyond?

We envision a dashboard system for decision support. We anticipate situations in which a U.S. analyst is tasked with monitoring all tweets posted by thousands of accounts already attributed to foreign adversaries. The analyst could pose a series of questions and use the dashboard system to find answers.

For example, starting with all tweets posted by the flagged adversary accounts, the analyst could query, "Show me all tweets that were posted in the last month, in the German language, from a supposed German location, and between 0200 and 0500 Central European Time," thinking that this is the period of night when most real Germans are asleep. The analyst could filter down further with a second query such as, "Show me all tweets that were related to the topic labels 'Coronavirus' or 'Social Distancing.'" If the analyst found a particularly interesting tweet, she could drill down even further, querying, "Show me all tweets like this one …" indicating that she wished to see all tweets associated with a similar combination of topics, "… from March 2020," remembering that this was the month in which Germany first imposed social distancing restrictions. Of these tweets, the analyst could then query, "Show me all tweets with the hashtag #Wuhan," and so on.

This human-driven, drill-down approach could allow the analyst to connect the dots between adversaries' tactics and strategies. Like in the query example we pose above, the analyst could use the dashboard to explore hypothesis about how adversaries sowed discord within NATO allies about the origin of the novel coronavirus during the early days of the pandemic (Rebello et al. 2020).

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
June 2021	FINAL	

4. TITLE AND SUBTITLE

Weaponized Tweets (Presentation)

5a. CONTRACT NUMBER

HQ0034-19-D-0001

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

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

5d. PROJECT NUMBER

CRP C2234

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Institute for Defense Analyses
Systems and Analyses Center
730 East Glebe Road
Alexandra, VA 22305-3086

8. PERFORMING ORGANIZATION REPORT NUMBER

IDA Document NS D-22805

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Institute for Defense Analyses
Systems and Analyses Center
730 East Glebe Road
Alexandra, VA 22305-3086

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.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Beginning in Summer 2018, the Institute for Defense Analyses (IDA) devoted limited internal research and development funds to exploring foreign malign influence operations in social media. To date, we have focused on the Twitter activity of the Russian-backed Internet Research Agency (IRA) in the run-up to the 2016 American Presidential election. We used a data-driven, quantitative approach with human-in-the-loop, employing unsupervised Machine Learning techniques (Latent Dirichlet Allocation) and other data analytic methods. We focused on the evolution of the adversary activity, tracking how topic patterns in that activity changed over time from 2012 through 2018. Our results uncovered multiple distinct tactical phases of the IRA attack. The IRA first began targeting Americans in late 2014 to early 2015. Their English tweet topics grew more specific over time, more varied, more negative, and more polarizing. Their final tweet topic pattern first emerged around late 2015, one full year before the 2016 election.

15. SUBJECT TERMS

Influence Operations; Internet Research Agency (IRA); Latent Dirichlet Allocation (LDA); Machine Learning; Russia; Social Media; Topic Model; tweet; Twitter

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Major, Philip L.
Unclassified	Unclassified	Unclassified	Same as Report	55	19b. TELEPHONE NUMBER *(Include area code)* 703-845-2201

Reset