INSTITUTE FOR DEFENSE ANALYSES

IDA Ideas (Podcast Transcript)—Weaponized Tweets: Artificial Intelligence to Defend Against Influence Operations in Social Media

Shelley M. Cazares
Emily M. Parrish
Jenny R. Holzer
Rhett A. Moeller

October 2020

Approved for public release;
distribution is unlimited

IDA Document NS D-17407
Log: H 20-000418

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

About This Publication

This work was conducted by the IDA Systems and Analyses Center under contract HQ0034-19-D-0001 under IDA's Independent Research Program, Project C2234, "Machine Learning for Social Media." The views, opinions, and findings should not be construed as representing the official position of the Department of Defense.

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

Executive Summary

IDA Ideas host Rhett Moeller spoke to researchers from the Science and Technology Division of IDA's Systems and Analyses Center about their use of machine learning to create a prototype system for analyzing Twitter posts that U.S. adversaries made to influence public opinion in the years leading up to the 2016 U.S. Presidential election. Joining him are Shelley Cazares, who leads the ongoing project, and two members of her team, Emily Parrish and Jenny Holzer.

The project began in 2018 with about three million tweets that had been posted from early 2012 through early 2018 by the Russian-backed Internet Research Agency. The team set out to create a prototype system using open-source software tools that could have helped intelligence analysts during that time. Using a machine learning technique called Latent Dirichlet Allocation and open-source software called MALLET, the team found that the topics of the adversary's tweets evolved over time into several tactical phases. Specifically, their English tweet topics grew tighter over time, more specific, more negative, and more polarizing, with their final pattern solidifying one full year before the 2016 election. With this system, this and other revealing information about U.S. adversaries' social media operations could be reported to U.S. Government decision makers with as little as one month of lag time.

Contents

The following transcript of "Weaponized Tweets: Artificial Intelligence to Defend Against Influence Operations in Social Media," an episode of the podcast IDA Ideas*, has been edited for clarity.*

Introduction

Rhett Moeller: Hello, listeners, I'm Rhett Moeller, and I'm the host of *IDA Ideas*, a podcast hosted by the Institute for Defense Analyses. You can find out more about us at www.ida.org. We also have a social media presence on Twitter and Instagram, so there are plenty of ways to keep up with the exciting work we're doing. Welcome to another episode of *IDA Ideas*.

Because of the ongoing COVID situation, we are conducting this episode by video conference, so there may be a slight difference in our quality. In this episode, we're going to take some time to talk about the interesting work going on at the Institute for Defense Analyses. Our research staff is driven by curiosity, a desire to better know and understand the world around us, and to find ways to use what we discover to help improve the safety of our Nation. Sometimes that work is directly tied to sponsor-driven requests, and sometimes it *anticipates* sponsor interest. Our topic today deals with this latter kind. There's a lot to cover, so let's get into it.

I'm joined by three of our researchers from IDA's Science and Technology Division: Shelley Cazares, Emily Parrish, and Jenny Holzer. Can you each take a moment to introduce yourselves?

Shelley Cazares: Thanks, Rhett. I'm Shelley Cazares. I studied electrical engineering and computer science at MIT [*Massachusetts Institute of Technology*] with my doctorate from [*the University of*] Oxford in the signal processing and neural networks lab. I spent a few years in industry before coming to IDA. Here at IDA, I focus on machine learning and computational modeling for defense in intel applications, especially in gray zone activities where influence operations fit right in.

Emily Parrish: Hi, Rhett. I'm Emily Parrish. Before starting here at IDA, I received my undergraduate degree in chemistry and computer science from the College of William & Mary and I'm currently working on my graduate degree at the George Washington University in data analytics. Specifically, I'm focusing on machine learning.

Jenny Holzer: Hi, Rhett. I'm Jenny Holzer. I came to IDA straight from academia after finishing my Ph.D. in physics at [*the University of*] Cincinnati, and then spending a few years as a postdoc with the Living States Physics Group at Vanderbilt [*University*]. Now at IDA I do mostly cybersecurity assessments of our Combatant Commands. So with a cyber background combined with being a regular social media user, I'm excited to join Shelley and Emily on this project.

Rhett: Thank you, and welcome to *IDA Ideas*!

About Machine Learning

Rhett: Shelley, you've really been putting your machine learning background to good use in leading an interesting project. Can you tell us about your work using artificial intelligence to analyze social media data?

Shelley: Well, artificial intelligence—it's such a broad term, and so is social media, so I will be slightly more specific. I'll say that we used machine learning to analyze adversary activity on Twitter.

A lot has been said about Russian interference in the 2016 election, using social media platforms like Twitter and Facebook. It turns out that many countries do this, and many of those countries do have adversarial relationships with the United States. Some countries started these campaigns well before the 2016 election, and some continue to this day. So platforms like Twitter and Facebook, they've all been working with law enforcement and the intelligence community to identify these "active measures" campaigns. Some platforms will then package up and release the posts made by the accounts, so that researchers like us can study them and learn from them.

I started an internal research project here at IDA back in summer of 2018, when Twitter first released the account names of over three thousand accounts posted by the IRA, which is the Internet Research Agency backed by the Russian government. In July of that year, 2018, Clemson University packaged up about three million tweets posted by those adversary accounts, and then they released them publicly on the FiveThirtyEight website, the data aggregation and polling website. I downloaded those tweets the very next day and the rest is history.

Now, since then, Twitter has released millions of additional tweets posted by thousands of additional accounts from many different countries, but my research actually started with the first dataset, which was posted by the Clemson academics.

As a machine learning researcher, I was so excited when I first learned about this dataset back in summer of 2018, because it was well constructed and it was well organized and it contained so much metadata. Each tweet was time stamped and date stamped and labeled by language, tagged with the country that it was supposedly posted from. So it was structured data, and it was available for almost-immediate analysis, which is a beautiful thing for a machine learning scientist.

Rhett: Absolutely, and as a librarian myself, I fully appreciate the beauty of structured datasets; it makes life so much easier. Many of our listeners who have worked in this area know what a huge field it is and that it continues to grow by the minute. And there are many ways to approach a project. What kind of machine learning techniques did you use in this project?

Shelley: We mostly used a technique called Latent Dirichlet Allocation, or just LDA. This is a form of unsupervised learning—it's a statistical model that automatically organizes the words in the tweets into underlying topics. So, a human user like you or me could then just glance through those topics to quickly get a sense of what our adversaries were tweeting about.

Supervised versus Unsupervised Learning

Rhett: Great, thank you. Now you used some terminology here—supervised, unsupervised. Can you take a moment to define what you mean by those?

Shelley: Sure. So there are really two main classes of machine learning techniques—supervised and then unsupervised learning. Supervised learning is when you have an answer key—you know what answer the system is supposed to spit out, and so you train the system by giving it data—training data, example data—and having it process that example data over and over again until it can spit out answers that are as close as possible to the answer key. Kind of like giving it old tests to practice on, over and over again, until you're pretty sure that it can get an A+ grade. Then once it's trained, you deploy it so that you send it out to take new tests on new data that you don't have the answer key for. So that's supervised learning. Now, unsupervised learning is different. With unsupervised learning, you don't have an answer key. Instead, you input the data into the system, and then you let the system figure out the best way to organize the data—the best way to rack and stack it—so that human users like you or me could understand it better. LDA is the technique that we used and that is a type of unsupervised learning.

Going back to LDA, it's a very well-known technique. It's been around for a while—I think the first paper was published in 2003. It's been coded up in several different open-source software kits that are freely available on the internet—we used the MALLET software. And we chose to use the software fresh out of the box, meaning that we deviated very little from its default parameters. We did that on purpose because that way our findings would outline the lower bound of capability—what kind of patterns, what kind of actionable intel that you could easily extract from the tweets. I guess more importantly, since LDA and MALLET were around back in the early days of Twitter, our work shows what kind of intel could have been extracted from these tweets back when our adversaries first got started in their social media influence campaigns.

Rhett: Thanks, Shelley. You did mention in your talk just now about some concepts and software that our listeners might be interested in knowing more about. We're going to do what we can to provide links to these topics and products like LDA and MALLET in our show notes. So listeners, if you're interested in learning more about these particular items be sure to check our show notes for direct links to more information.

Beyond Troll Detecting

Rhett: Now, Shelley has alluded to the global concern over this disturbing trend. Jenny, you took a look at what other research groups were doing in this area. Are there other research groups using this same technique?

Jenny: Yeah, so one of the first things we did was review the literature, and we found three main things.

As soon as this data was released to the public in the summer of 2018, researchers from around the world from all different disciplines just jumped on it. These researchers were from data science, social science, international relations, you name it. And that's actually the whole reason that the Clemson University researchers shared their dataset. They actually said that they wanted many more brains looking at it to see what else they could find in the data. The findings started coming out pretty immediately, within the first week. Some of those focused on visualizing networks of Twitter users or focused on the overall topics that the trolls tweeted about. Journal articles from groups of researchers applying machine learning techniques to analyze the data started coming out just a bit later.

Which brings up the second point that most groups who were using machine learning were building troll detectors. These troll detectors are systems that could take in a tweet and determine if it was written by a regular Twitter user like you or me, or a foreign troll.

Rhett: Okay, hold on a second. I'm getting flashes of fairy tales here, what is a troll, Jenny? Can you explain what that is?

Jenny: Yeah, of course. Trolls on social media are people who deliberately try to provoke others by saying inflammatory or offensive things. If you've spent any time on social media at all, you have absolutely encountered these trolls. Some of them are just doing it just for fun, for the fun of riling people up. Other trolls are really trying to create a more calculated response, as we'll talk about later.

But now getting back to the literature review, the third thing we found was that most of the early studies attempting to understand the tactics and the strategies of trolls often ignored temporal information. I actually recall one article from a group of researchers in Australia. In their article, they called this out as a limitation. Their article said that if temporal analysis was done, it typically lumped all the tweets together, and then looked at the number of tweets or the frequency of a particular hashtag at different points in time, rather than looking at how the topics that they tweeted about changed over time.

Rhett: Okay, I can see how that would be a problem, definitely, and it's a problem that we are seeing around the world, this troll interference and so forth. The potential for harm has certainly captured the attention of researchers. And so, Shelley, can you tell us how this initial literature review informed your work?

Shelley: It was really helpful in getting us started. First, we did not want to make a troll detector, like many of the other groups were doing. To do that, you would need to use supervised learning techniques where you have an answer key to a training set of tweets. You know whether the tweets in your training set were posted by a real foreign adversary troll or a real regular person. To do that, you need two sources of data—tweets from real trolls, which we did have, thanks to Clemson and FiveThirtyEight, but then also a second set of tweets from real regular people, which we did not choose to harvest. So, that precluded us from developing a foreign troll detector. And besides, trolls are not unique to foreign adversaries—we grow them ourselves here in the U.S. So we did question just how useful a foreign troll detector would really be in the great scheme of things.

Second, we didn't want to lump all the tweets together over time because that lumping together is retrospective in nature. It's easy to pick out the patterns in the tweets when you have 20/20 hindsight—when you can see all the tweets at once. We wanted to put ourselves into the mind of an intel analyst, back in the early 2012 time frame, when the Clemson data first started. And so we wanted to base our analysis on a sort of thought experiment.

Thinking it Through

Rhett: Okay, great! And just for the sake of illustration, can you walk us through this thought experiment?

Shelley: Sure. Picture this—Washington, DC, 2012. Hypothetically speaking—I'm just making this up—let's say that a U.S. intel organization gets a high-confidence tip that an adversary nation state is clandestinely posting tweets on social media. And let's say that the source even identifies the particular Twitter accounts—those that, years later, Twitter will eventually attribute to the IRA. But back then, in 2012, imagine that an intel analyst is tasked to monitor those flagged accounts for indications and warnings that our adversaries are using social media maybe even to launch an information warfare campaign against the United States. What does that analyst do? If you were the analyst, what would you do? What information would that analyst have known at the time? What tools would she have had available? You know, the LDA technique was alive and kicking back then, so she could have used that. But *how* would she have used it? And *how* would she have analyzed the results? And what would she have *done* with those results in real time?

That's the thought experiment that really drove our analysis. We set out to create a proof-of-concept prototype system using open-source software tools—like LDA and MALLET—that could have helped that analyst practice her tradecraft. Our system was not intended to replace that analyst, but rather help that analyst perform the same type of tradecraft she always does, and apply it to millions of tweets instead of just dozens of documents.

Rhett: Got it, and this raises a whole bunch of questions like how this hypothetical intel analyst would have used your system. Could she have just poured tweets into it, or pushed a button and sat back and let the system do all the work? Or was there some other pre-processing that she would need to do in order to get it ready for this? Emily, can you help explain this?

Emily: Sure, Rhett. About 90 percent of a data scientist's, or specifically a machine learning scientist's, job is actually those pre-processing steps. Things like data cleaning, data normalization, putting the data in a form that the machine learning model can actually crunch through it. We were really fortunate that the data that we were dealing with was very well-structured, so the main things we really had to take care of were the cleaning and the reformatting parts. First, what we did was we binned all the tweets, which were parsed into individual files by month. This was just done so we could look through the evolution of the topic patterns from month to month, like Shelley said. We also separated out the English tweets from all the others, since we had hypothesized that the English tweets were posted with this American or Western audience in mind.

Ultimately we had to address the big technical question here, which was: Can we use machine learning techniques like LDA on such short pieces of text like tweets? Usually LDA is applied to long passages of text—entire documents or long reports. So could it cluster the topics with the documents of only 280 characters or even less? We did a few experiments to find out how short a text file we could really process with LDA. And what kinds of characters we needed to filter out before applying LDA—punctuation, emojis, hashtags, or even URLs. In some cases, we found it was best to remove some of those elements before inputting the tweet into LDA, and in other cases, we found that it was important to keep it in.

Rhett: Okay, very interesting. Now you described what the hypothetical intel analyst would have done to input the tweets into your system. Presumably this could mean thousands, maybe even millions, of individual items—so what does that mean for the system's output? Shelley?

Shelley: Well, remember—LDA is an example of unsupervised learning, so that means it doesn't require an answer key in order to learn from the data. That's great, because it could take time and money to compile an answer key. You'd have to get humans to read through thousands of example tweets and label them just so. But we didn't have to do any of that, because LDA doesn't need an answer key. Instead, it lets the words in each month's worth of tweets organize themselves, with no answer key for guidance. So, no human input was needed up front. But, there's a catch—there's always a catch. You can't get something for nothing. Unsupervised techniques like LDA—they may not require human input *up front*, but they do require human input at the *end*. So that's where Emily came back in. She didn't need to label the individual tweets, but she did need to label the tweet topics. Although there were a few *million* tweets, there were only a few *hundred* topics.

Rhett: That's a significant difference. And it makes sense that work has to come in somewhere—there's always work to be done. Emily, can you explain your process? What is a tweet topic? What was your approach?

Emily: Sure, Rhett. So the main outputs of LDA model are the topics. If we are looking at a particular month bin, each topic is just a list of words from that month's worth of tweets. At a high level, these words will cluster because the words seem to appear together a lot, in the same tweets, over and over again. So the LDA model learns these clusters and groups them together into what we call a topic. Just to provide an example, a topic could be a list of words like *sports*, *nfl*, *super*, *game*, *bowl*, etc., and when we put some semantic context to it, it's clearly a sports-related topic, so the label would simply be "sports." But LDA, as we see in this example, doesn't really provide a semantic meaning for why these words are grouped together, it's just seemingly a random list of words, and that human needs to provide insight and I served as that human for our analysis. If we reference back to that thought experiment that Shelley discussed, I kind of stepped into that role of the hypothetical intel analyst.

Rhett: Well lucky you, Emily. Sounds like you had a very important part in the processing. Why did Emily get to play this role? How did she get to be the one to provide that human input? Jenny, do you have some insights into that?

Jenny: Yeah, of course. To provide the semantic labeling, or the *why* behind the automated word groupings, you need a human expert. By "expert" here, I mean someone who understands the nature of the source and the context of the data. This would be true for any machine learning project. But here, for this project, we're talking about tweet data, and so we needed someone who understood Twitter. We needed someone fluent in English, since English was the most predominant language of the tweets. We needed someone with a strong cultural understanding of the trends and the memes and the hashtags used in American online culture, since most of these tweets were spoofed to make it look like they were actually posted from a Twitter user in the U.S. So, again, we needed someone who spoke English, understood American cultural memes, and had a strong familiarity with Twitter, and Emily fit that bill.

Rhett: Great. So Emily, what did you do in this role, specifically?

Emily: My task was to essentially look at each grouping of words, or each topic, and assign a one- to two-word label to that topic, and that was to semantically summarize what the topic was about. We set the LDA model to fit ten topics per month. There were over thirty months in the dataset, so that meant over three hundred topics that I had to label. It was important that I didn't bias myself in that process—that I didn't, say, focus too heavily on one month at the expense of other months. So, I wanted to print them all out. We removed all the furniture in Shelley's office and I actually spread them all out of the floor, so I could see them all at once and together. Then I just got down on my hands and knees and I read all of the topics going forward in time with my human eyes. In that process, I marked them up using different color highlighter pens so that I could track if each given topic seemed to carry on from month to month. It was super crucial that I looked at the months like this because sometimes the topics would disappear and then reemerge a few months later. And then some of the topics were actually cyclical, like the Christmas topics popped up pretty much every December.

Rhett: Oh, that's really interesting. So it sounds like there was a lot of this spreading out on the floor and looking at it with your human eyes. How long did it take you to label that many different topics?

Emily: It took about a full workday, but knowing what we know now, we've actually developed some more high-tech visualization techniques and a better SOP [*standard operating procedure*] that probably could speed things up.

Patterns Emerge

Rhett: Interesting, and obviously going through you began to see different patterns emerge. Were there any surprises in this part of your work? I imagine that with so much to work with, you were bound to find something. Shelley, do you have something on this?

Shelley: Yes, we found that our adversary's topics did not remain constant. 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. So to be specific, their English tweet topics grew tighter over time, more specific over time, more negative, and more polarizing. Their final pattern solidified around fall 2015, which happened to be one full year before the 2016 election.

Rhett: Interesting—a progression toward a more defined pattern would certainly seem to indicate intent.

Shelley: Sure. So, the first tactical phase we found was 2012 to 2014. The earliest tweet in the dataset was February 2012, so this was the first two years of the dataset. Back then, only a few dozen tweets were posted each month. The system could have gotten by with no machine learning at all—so no topics to spit out. Just histograms of statistics—bar graphs to show that most tweets were in Russian, with a few in English. Our hypothetical intel analyst could have looked at each tweet individually—and she would have just mostly seen Cyrillic characters. So I think that she would have surmised that most of these tweets were not meant for an English-speaking audience in the U.S. She probably would have just passed these tweets off to the appropriate language or regional expert and then settled back into a sit and watch mode for the next two years.

Rhett: A lot can happen in two years! Did you see any definite changes during that period?

Emily: Not really, no. We saw things change *after* that, in late 2014 to early 2015. At that point, the adversary began tweeting more frequently, by over two orders of magnitude. And by January 2015, the most prominent language was now by far English. So our analyst could have at this point sat up and noticed a change in behavior of the data itself. And then, of course, to characterize that behavior, she wouldn't have been able to read through all the tweets since there were tens of thousands of them in just one month. But, that's where our system could have come in. It would have made her workflow easier and automatically organized the main topics of the tweets for her to label.

I guess, for instance, we found that many of the English tweets in January 2015 used words like: *make*, *love*, *time*, *thing*, *success*, *give*, *smile*—this loose, vague, and feel-good positive topic that our analyst could have labeled just "motivation" for lack of a better word. Also many tweets used words like *news*, *local*, *police*, *fire*, *killed*, and so on, and that's another loose, vague topic, but this time it's kind of negative, so she could have labeled it "local news."

And then also, we go back to our previous example [*with*] words like *sports*, *nfl*, *supe*r, *game, bowl*—remember, this was *January* of 2015. It's obvious this was a sports topic, but probably more specifically the Super Bowl.

So based on all these topics—most really vague and loose, and some with positive and some with negative affect, we suspect that the intel analyst, in real time back in 2015, could have been pretty unsure about what was going on based on the pattern of topics.

Rhett: I presume this phase could then be seen as baiting the hook, right? Using mainly benign themes to draw the largest possible crowd, but then setting the stage for the real purpose with some of these heavier words. Jenny, do you have any insights into this?

Jenny: Yeah, then we saw the pattern change just a few months later. In the summer of 2015, the adversary started tweeting about a single cultural topic. Emily determined this to be exercise. But she still saw that a little bit of news and a little bit of politics was still thrown in the mix. They also started tagging specific Twitter users. We think they were trying to do this to build a following at this time. But at this time, our hypothetical intel analyst would probably still be confused about what the adversary's underlying strategy was.

Rhett: Sure, it's really all over the place. Well, considering again how scattered the approach was, I can see why it would be confusing to somebody at the time. Between the sheer volume of information as well as its scatterplot nature, there sure is a lot to take in. Shelley?

Shelley: And then it all fell into place a few months after that. In fall of 2015, the topics were suddenly tighter, more specific, more varied, and more polarizing. I mean, many tweets were still associated with the local news topic and the politics topic. But now many more of the politics tweets had adopted a more polarizing tone—less like "this candidate is running for president" and more like "this party will slit their wrists if this candidate wins the primary." And there were also some really polarizing topics that emerged, like guns and terrorism. This is the time frame when things got very interesting. Fall 2015—one year before the 2016 election—that was the earliest point in time in which our proof-of-concept system could have helped reveal what we now all know was our adversary's strategy—to sow discord, especially in the English-speaking world.

Rhett: Wow, that is a significant shift in message! Did their tactics change again after that, Emily?

Emily: Based on our analysis, their tactics stayed pretty much the same after that. For example, in October of 2017, we saw specific, polarizing topics like the NFL national anthem protests, Hurricane Maria in Puerto Rico, and the Las Vegas shooting.

Shelley: Yeah, and many of those later topics in 2016 and 2017, they continued on, like those relating to specific topics in news and sports and politics. Also many later topics related to race relations, which is a very polarizing issue. I guess you could say, from the perspective of this adversary, the way to Americans' hearts is news, sports, politics, and racism.

Rhett: Wow.

Jenny: But this dataset ended in spring 2018. And from our ongoing research, it doesn't look like that much has changed since then. The adversary's patterns seem to have remained the same.

Four Takeaways

Rhett: Okay, great. Thank you so much for sharing your analytical experience! Now for the "so what." What does all this mean?

Shelley: Well, there are really four takeaways here, and I think we could each take one and then I could wrap it up. First, just because our adversary's tactics haven't changed that much in the last couple of years doesn't mean that they won't change a lot in the future. They did change a lot in just 2015 alone—from those loose and vague motivation and news topics in early 2015, to the single-topic exercise tweets in mid-2015, to those specific and varied and polarizing topics in late 2015. So, we know our adversaries can adapt because they have adapted. And so the United States needs to be able to adapt, too. Throughout our thought experiment, we kept envisioning two people in our heads—the hypothetical intel analyst, like we've talked about already. But also the troll manager—tucked away in his or her office, thousands of miles away from the United States. What was he or she directing their staff to tweet about each month? What was written on their whiteboard? When did they erase their whiteboard and start all over again? And did their strategy remain constant over time and it was just their tactics that changed, or did their underlying strategy change over time, too? To get at those kinds of questions in the future, the U.S. Government will need machine learning systems that can adapt over time, if, and let's face it, when our adversaries adapt.

Rhett: Absolutely. Understanding adversary mindset is obviously a huge advantage, for sure. But definitely keeping in mind the analyst is an important consideration; if the system is too unwieldy, it's going to fall out of favor quickly, especially as the rate this information flow increases.

Emily: Yes, that's actually our second takeaway point. Any machine learning system has to scale. During the months with the most tweets, the pipeline handled tens of thousands of tweets and could easily have handled more. But as the human analyst, I was the bottleneck in terms of scalability. I had to read through the topic words for each month, and apply those semantic labels to them. That was tricky and it took some time. Since then, you know, we've worked out an SOP for how to make that easier on the human, but it still takes discipline and time for that human labeler. Also, if we wanted to do that labeling on a week-by-week process or a day-by-day basis, could a human keep up with that pace? Probably not with the right-out-of-the-box LDA techniques that we used. Newer machine learning techniques could really track the tweet topics from month to month and that would automate some of that labeling process to lighten the load on the analyst. Then maybe she could respond more quickly to more tweets.

Jenny: Also, a third thing we took away is that this kind of influence operation could happen at any time. Not just in the run up to an election like we've been talking about, but for any event, like a pandemic or like a protest. A lot's been written about all the disinformation surrounding the protests in Hong Kong last year and more lately about the coronavirus pandemic. So, for the future, we're envisioning sort of a dashboard system for decision support. Any time an analyst is tasked with monitoring accounts that have already been attributed to foreign adversaries, this analyst could pose a series of queries. It might go something like this: For all tweets posted by the adversary accounts, show me all that were posted in the last month, in the German language, from a supposed German account, and between 2 a.m. and 5 a.m. central European time. Because, of course, you would think that most real Germans should actually be asleep at that time. And then let's filter further: Of those results, show me all that were related to the topic labels "coronavirus" or "social distancing." And, hey, this looks like an interesting tweet. Show me all the tweets like this one—with a similar combination of topics and show me all the ones dating from March 2020 because when Germany first imposed social distancing restrictions, this date's important. And then of these, show me all tweets that included the hashtag Wuhan. And so on and so on. This type of human driven, drill-down approach lets that analyst explore her hypotheses about how the adversary sowed discord among NATO allies about the origins of the coronavirus. That's just one example.

Rhett: I can definitely see that sort of dashboard setup being extremely useful, especially for somebody who's in the business of detecting threats!

Shelley: And our final take-away is: This is a multidisciplinary problem. So you really need experts in three things: First skillset is science and technology to create those advanced systems that can automate more of the topic labeling process, leaving less load on the human analyst, as Emily talked about. That's what we're working on right now here at IDA. The second skillset is geopolitics and foreign language to understand what events are happening in different regions of the world, as Jenny talked about. And our former military intelligence colleagues here at IDA have been helping us with that. The third skillset is social science. Look, I'm a machine learning scientist, so I focus on data, data, data and algorithms, algorithms, algorithms. But I will be the first to say that 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 act on tweets, all against the backdrop of their own cultural experiences. Rigorous social science research is needed—not only to help analysts understand the answers that their systems spit out in response to their queries, but to help them formulate their queries in the first place. And my social scientist colleagues here at IDA have been helping me better understand that over the last year or so.

Closing

Shelley: The good news, really, is that the U.S. really does have all of these skills sets in abundance—science and technology, geopolitics and foreign language, and social science. The hard part really is just getting all of those experts together under one roof and herding the cats.

Rhett: Absolutely, and that's not just the U.S.—IDA certainly has a lot of experience in each of these areas. We are definitely interested in doing more work in this area, and we have a lot to contribute to national security when it comes to this sort of analysis.

Shelley, Emily, Jenny, thank you very much for taking the time to discuss this intriguing project with us and for giving us more insight into an interesting yet serious topic. It has been most illuminating!

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.

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*
XX-10-2020

2. REPORT TYPE
Final

3. DATES COVERED *(From - To)*
Aug 2020 - Oct 2020

4. TITLE AND SUBTITLE
IDA Ideas (Podcast Transcript)—Weaponized Tweets: Artificial Intelligence to Defend Against Influence Operations in Social Media

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.
Moeller, Rhett R.

5d. PROJECT NUMBER
C2234

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
IDA Systems and Analyses Center
4850 Mark Center Drive
Alexandria, VA 22311-1882

8. PERFORMING ORGANIZATION REPORT NUMBER
IDA Document NS D-17407

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
IDA Systems and Analyses Center
4850 Mark Center Drive
Alexandria, VA 22311-1882

10. SPONSOR/MONITOR'S ACRONYM(S)
IDA SAC

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This document is an edited transcript of an episode of a podcast called *IDA Ideas*. In this episode, host Rhett Moeller interviews researchers from the Science and Technology Division of IDA's Systems and Analyses Center about their use of machine learning techniques to analyze use of Twitter by U.S. adversaries to influence public opinion of current events. Joining him are Shelley Cazares, who leads the ongoing project, and two members of her team, Jenny Holzer and Emily Parrish. Using a machine learning technique called Latent Dirichlet Allocation and open-source software called MALLET, the team found that the topics of the adversary's tweets evolved over time into several tactical phases. With their prototype system, information about U.S. adversaries' social media operations could be reported to U.S. Government decision makers with as little as one month of lag time.

15. SUBJECT TERMS
machine learning; 2016 presidential election; Russian tweets; active measures; influence operation; information operation; Twitter; social media; MALLET; Latent Dirichlet Allocation (LDA); topic model; unsupervised learning; supervised learning

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Buckley, Leonard J.
Unclassified	Unclassified	Unclassified	Same as Report	16	**19b. TELEPHONE NUMBER** *(Include area code)* (703) 578-2800