Detecting the Spread Misinformation in Social Media

MARY E. NWOSU*, Department of Electrical Engineering and Computer Science - Howard University, USA

Fig. 1. The Applied Research Laboratory for Intelligence and Security (ARLIS)

In the summer of 2021, The Applied Research Laboratory for Intelligence and Security (ARLIS) [a University-Affiliated Research Center (UARC) based at the University of Maryland College Park (UMD)] hosted the virtual Research for Intelligence Security Challenges (RISC) Initiative internship program. This 10-week summer program for hard security problems paired students with mentors from the UMD campus and the Department of Defense (DOD) and the Intelligence Community (IC) community. As a RISC summer intern, I was assigned with **building on existing open source code to create a multi-platform social media simulator that can support human subjects experiments on information spread**. I continued my participation with this assignment during the 2021 Fall Semester as a Research Assistant.

The aim of this report is to detail the efforts of that RISC internship and the Research Assistantship to date in the following 7 sections: Introduction and Motivation, Literature Review, Proposed Approach, The Simulation Model: Discovery, The Pilot Study, Performance Evaluation, and Conclusion.

CCS Concepts: • **Artificial Intelligence** → **Machine Learning**; • **Machine Learning Techniques** → *Detection and Mitigation*; • **AI-ML Techniques** → Detection Algorithms; • **Neural Networks** → Model Accuracy.

Additional Key Words and Phrases: artificial intelligence, machine learning, misinformation, disinformation, neural networks, detection, mitigation, social media, model accuracy, stance detection, fake news, automated fact-checking, sentiment analysis, Twitter data, Facebook data, TikTok data, COVID-19, Zika, disease vaccinations

ACM Reference Format:

Mary E. Nwosu. 2021. Detecting the Spread Misinformation in Social Media. In *EECE 680: Reading Research, 2021 Fall Semester, Howard University*. ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Manuscript submitted to ACM

1 INTRODUCTION AND MOTIVATION

1.1 ARLIS: The Applied Research Laboratory for Intelligence and Security

The Applied Research Laboratory for Intelligence and Security (ARLIS), based at the University of Maryland College Park, was established in 2018 under the sponsorship of the Office of the Under Secretary of Defense for Intelligence and Security. As a University-Affiliated Research Center (UARC), ARLIS' purpose is to be a long-term strategic asset for research and development in artificial intelligence, information engineering, and human systems. By developing solutions that effectively combine humans and technology, rather than focusing on humans or technology, ARLIS ultimately seeks to have a significant and positive impact on our lives, our communities, our nation, and the world.

ARLIS serves the public interest as independent technical leaders with a clear purpose of building and maintaining a long-term, committed relationship with the DoD and its partners. It's mission is to help to assure the U.S. and allied decision-advantage in long-term Human Domain competitions, in part by leveraging our core competencies in social and behavioral sciences, data science and infrastructure, AI and machine-learning (ML), engineering, and advanced computing. Specifically, ARLIS has seven core mission areas [1] Cognitive Security, [2] Acquisition Industrial Security, [3] Modeling Mitigating Insider Risk, [4] Augment Collective Intelligence, [5] Language Culture, [6] Artificial Intelligence, Autonomy, Augmentation, and [7] Human Performance.

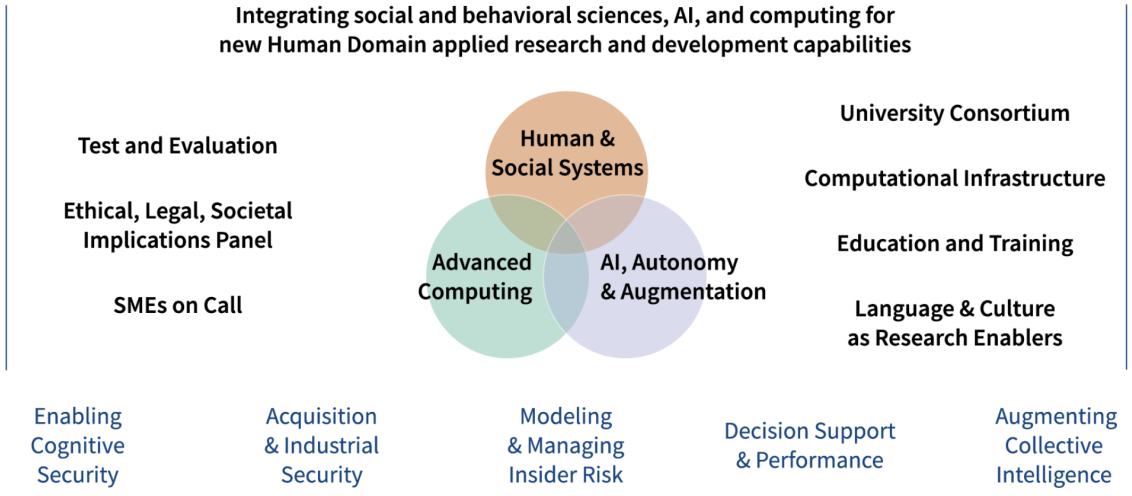

Fig. 2. Core Mission Areas of ARLIS and How They Fit Together

1.2 Cognitive Security (COGSEC)

Disinformation is one of the most critical issues of our time, concerned with online and offline influence at scales ranging from individuals to large populations. Operations in the Information Environment (OIE) are conducted within the context of Cognitive Security (or COGSEC). The movement toward symbiotic human-machine interfaces creates an urgent demand for research to inform operations in the broadest sense.

The ARLIS COGSEC program is developing both targeted projects and overarching capabilities ready to use for a broad range of research, wargaming, and operational questions, with goals including the following:

- Design online systems and interactions to reduce vulnerability to misinformation and manipulation;
- Detect and mitigate targeted information manipulation attempts targeted at governmental insiders; and
- Develop the integration of cyber and social media systems or simulations while also monitoring factors outside social media environments.

Adversarial entities worldwide continue to spread disinformation on social media and have revealed a severe vulnerability in the United States' security and its Western allies. Several projects, funded through the COGSEC mission area, investigate the spread of information campaigns by examining how different emotions influence resharing content in Polish and Lithuanian socio-political social media.

COGSEC's multinational DoD Minerva Research Initiative program collects and analyses real-world Facebook and YouTube data from Poland and Lithuania - countries that were chosen for their strategic relevance to NATO and Europe. Researchers annotate samples of over 1000 public Facebook posts and 300 YouTube videos from each country for emotions and topic content. COGSEC also oversees computational linguistic analyses from 2015 to 2020 to examine sociopolitical topics and cross-platform information spread. Most importantly, COGSEC programs, like the Minerva project, address critical gaps in research about how information spreads across social networks. If successful, researchers can enhance understanding of how emotion can affect behavior online and what types of emotional content are most likely to make messages go viral, for good or ill.

1.3 RISC: Research for Intelligence and Security Challenges

In the summer of 2021, ARLIS hosted the virtual Research for Intelligence Security Challenges (RISC) Initiative internship program, building off the success of the 2020 AI Research for IC Challenges (AIRICC) program. This 10-week summer program for hard security problems paired students with mentors from the UMD campus and the Department of Defense (DOD) and the Intelligence Community (IC) community.

Students were introduced to career opportunities with the DoD and IC as well as develop their technical capabilities through hands-on real-world problems, posed by government operators and supported with realistic data sets and other materials. Over the 10-week period (June 1, 2021 to August 6, 2021) students participated in lectures and regular team development meetings in a shared virtual work environment. The summer program concluded with a demonstration event and workshop, with a panel of visiting experts from DOD/IC to discuss the results.

ARLIS' core mission areas supported in Summer 2021 included Cognitive Security (COGSEC), Artificial Intelligence, Autonomy, and Augmentation, Modeling Mitigating Insider Risk, and Acquisition and Industrial Security, as well as geospatial analysis, human geography, and critical technology protection. These mission areas were supported by interns with expertise in one or more of the following disciplines:

- **Computer Science, Information Science and Engineering**: AI/ML algorithmic development, HCI, data science, data and knowledge engineering, software engineering, systems engineering, media analysis and forensics, information systems design, GIS;
- **Mathematics and Statistics**: Data analytics, quantitative modeling, experimental design, graph analytics, data visualization;
- **Social and Behavioral Sciences**: anthropology, human geography (e.g., pattern of life and mobility modeling), cognitive/neuroscience and psychology, criminal justice, teamwork, and group dynamics, communications, disinformation, and misinformation;
- **Library Science**: Data curation, tagging, metadata, repositories, social media analytics;

- Additional topics included: Measurement and evaluation of learning outcomes, environmental modeling and remote sensing, human factors, regulatory public policy.

1.4 Topic N: Cognitive Security Social Media Simulation

As a 2021 summer RISC intern, I was assigned to "Team N" - which consisted of me, Mary Nwosu (a 1st year PhD student in Computer Science, Howard University) and Evan Jones (a 4th year PhD student in Government Politics, University of Maryland) - to support the "Cognitive Security Social Media Simulation" topic. Our team was sponsored by the Office of the Under Secretary of Defense for Intelligence and Security, or OUSD(IS), under the mission area of Cognitive Security (COGSEC), and the team was directed by our faculty mentor, Dr. Ruthanna Gordon. I continued to participate as a member of this team during the 2021 fall semester as a Research Assistant.

COGSEC is concerned with online and offline influence — and protection from influence — at scales ranging from individuals to large societies. Information conflict is one of the most pressing issues facing the United States, including not only formal Information Operations (IO), but a wide range of military and civilian sources—and targets—of misleading and harmful messaging. ARLIS's interdisciplinary human domain experts are bringing together experimentation and cutting-edge computational methods to enable US advantage in this area, and to restore and defend an information environment that facilitates robust, well-informed democratic discourse and institutions.

As part of this effort, Team N's assignment was to: **Build on existing open source code to create a multi-platform social media simulator that can support human subjects experiments on information spread.**

This report details the efforts of Team N in the following 5 sections: The Literature Review, Proposed Approach, The Simulation Model of "Discovery", The Pilot Study, Performance Evaluation, and Conclusion.

2 LITERATURE REVIEW

2.1 The Spread of Information as Influence

Cognitive Security (COGSEC) is concerned with online and offline influence – and protection from influence – at scales ranging from individual to whole-of-society. Adversaries, competitors, partners, and friends of the U.S. are all engaged in influence campaigns, which are growing in scope and scale. ARLIS's interdisciplinary Human Domain experts are bringing together applied experimentation and cutting-edge computational methods to enable U.S. advantage in this area, and to restore and defend an information environment that facilitates robust, well-informed democratic discourse and institutions. However, ensuring consistent U.S. advantage in this domain requires us to understand and address these problems quickly, effectively, and consistently. ARLIS has been working to expand and improve on preexisting applied research methods to achieve these goals, at scale, in an increasingly complex social and informational environment.

Imagine that an adversary has dramatically increased their information operations in advance of an election in a U.S. allied nation. An effective U.S. response requires rapid sense-making and intervention capabilities, but also an extensive foundation of previous study, modeling, and analysis to ensure that the sense-making is accurate and the interventions effective. These needed capabilities include:

- Research for understanding the information environment. High-fidelity live, virtual, and constructive studies can help prepare us for urgent IO needs by answering questions about how information and disinformation spread between platforms and networks, what makes populations vulnerable or resilient to influence, and what influence strategies are most effective for what goals;

- Test and evaluation of operational tools. Vendors and developers offer a seemingly endless supply of tools claiming impressive operational abilities. Independent evaluation in realistically-simulated IE (Information Environment) settings can provide clear, quantitative evidence for the efficacy of these tools, allowing the best to be ready for quick-turnaround deployment when needed; and
- Wargaming. Including high-fidelity IE simulations as a component of wargames allows participants to try out new strategies and practice decision making that reflects and applies to future operational work, replacing "white card" IE representations (i.e., descriptions of events requiring minimal in-game response) that place few demands on Blue team strategy.

As adversarial entities worldwide continue to spread misinformation on social media, they have revealed a severe vulnerability in the United States' security. COGSEC address these critical gaps in research about how misinformation spreads across social networks.

2.2 The Spread of Misinformation

The spread of misinformation is one of the most pressing issues facing the United States today, not only in formal governmental Information Operations, but also in a wide range of military and civilian sources. Misinformation is false, inaccurate, or misleading information and messaging that is communicated regardless of an intention to deceive (i.e., false rumors, insults, pranks, and misleading use of facts). Disinformation [19] is considered a subset of misinformation that is deliberately deceptive. Although, the efforts of Team N primarily focused on misleading information and messaging, regardless of intentionality, it is worth noting again that U.S. adversaries, competitors, partners, and friends are all engaged in disinformation activities and these activities are growing in scope and scale. Fake news - defined by Cambridge Dictionary as "false stories that appear to be news, spread on the internet or using other media, usually created to influence political views or as a joke" [9] - is a prime example of disinformation. And the spread of fake news is currently and profoundly affecting many aspects of U.S. society by creating biased opinions and manipulating mindsets [19].

2.2.1 The Spread of Misinformation in Social Media. With the advent of the Internet, the rapid adoption of online social media platforms quickly followed. And as society is beginning to feel the collective toll of the spread of misinformation; nowhere is this evidenced more so than in the realm of these social media platforms. Facebook, Twitter, Instagram, TikTok, and other social media networks have flourished in the past decade and have increasingly influenced people's daily life and behaviors. With access to vasts amount of data that is shared on these platforms, people consume information more readily and effectively than through traditional media (newspapers, network television, cable news programming, or magazine articles) or through other methods of information retrieval (online search engines, library catalogs,...). While social media platforms allow users to discuss, share news, and form groups around issues like democracy, education, and health; such platforms are simultaneously used to spread misinformation - intentional or not. The dissemination of misinformation, exacerbated by the very nature of online social media platforms, can lead to detrimental outcomes - most notably in U.S. national security.

2.2.2 Public Health Threats. Now more than ever - in this time of the COVID-19 pandemic - threats to public health of the U.S. arise from the spread of misinformation [1, 5, 17, 21, 33]. This is particularly true in areas such as vaccination hesitancy in vulnerable populations [31] as well as in general sentiments towards other health topics [11, 22, 39].

2.2.3 Domestic Terrorism Threats. The online proliferation of "The Big Lie" is the fake news at the center of the insurrection at the U.S. Capitol on January 6, 2021. This disinformation fueled hundreds of violent extremists to disrupt the joint session of Congress assembled to count electoral votes. The "Stop the Steal" campaign (a manifestation of "The Big Lie") not only rejected the legitimate results of the U.S. 2020 Presidential election [7, 16, 23, 24, 34], but also fed into the political radicalization of American citizenry – and left, ultimately, 9 dead, 138 police officers wounded, and many more injured. What if the spread of this fake news within online social media platforms could have been detected beforehand? Could the 2021 insurrection at the U.S. Capitol have been prevented?

With detailed guidance under development regarding required capabilities to address these threats, ARLIS's COGSEC mission area will integrate live experimentation, quantitative and qualitative modeling, and well-grounded simulations that facilitate these new methods in order to provide U.S. intelligence and security operations with evidence-based advantages.

3 PROPOSED APPROACH

3.1 The Goal of COGSEC

The ultimate goal of ARLIS's COGSEC mission area is to gain insight into information environment problems, and research and build potential solutions, with a complete research pipeline that takes ideas from basic research to field readiness in a way that is operationally-driven, timely, and trustworthy. In service of this goal, COGSEC is developing a Cognitive Security Proving Ground (CSPG) into a comprehensive environment for scientific study, test/evaluation, training, wargaming, and mission planning of information and influence operations. The Proving Ground will reduce risk on Information Operation (IO) solutions prior to use in the wild, offering the capability to test solutions in a controlled, secure, and realistic environment. ARLIS's capabilities will:

- Combine human and automated actors in a simulated messaging environment where we can confidently track exposure to information, interactions around that information, and responses to a range of messaging;
- Measure the effectiveness and potential unintended consequences of proposed capabilities;
- Measure the effectiveness and potential unintended consequences of proposed capabilities; and
- Create metrics for detecting, understanding the impact of, and countering key adversarial narratives in social media streams.

To support these capabilities, ARLIS will bring together:

- Researchers, operators, and problem solvers in social and behavioral sciences, computational analysis, and related fields;
- A dynamic library of tested software-based tools and models, along with rigorous methods for evaluating and adding new tools and models; and
- Simulation environments that accurately represent the complexity of real-world IEs.

Along with these technical capabilities, we are developing new applied research methods to leverage the potential of LVC (Live, Virtual, and Constructive) simulation and increase our ability to characterize and create effective interventions and defenses for information and influence.

3.2 The Pilot Study and Team N

The activities described in this report include initial proof-of-concept pilot work on both the research methods described above and the infrastructure to support them. They are intended to feed into both immediate and long-term plans for building out the full Cognitive Security Proving Ground and meeting U.S. government needs for Cognitive Security.

Pilot activities are small-scale prototypes for capabilities and methods that will ultimately be used to address urgent customer Information Operation and COGSEC needs. The goal is to identify potential problems early and at minimal cost, develop mitigation strategies to address those challenges in future work, and exercise simplified versions of planned functionality.

For the current task, Team N developed a beta version of a social media simulation platform instrumented for research purposes. While a number of social media simulation platforms exist elsewhere, they suffer from issues on several fronts:

- Access: Platforms are not consistently available in a timely fashion for ARLIS customer needs;
- Simulation complexity: Interaction with messaging is artificially formatted (e.g., users see only one message at a time rather than a more realistic social media feed) and interaction between users is artificially constrained (e.g., no live social interaction is possible); and
- Instrumentation: Many high-fidelity messaging simulations are designed for training rather than research, meaning that while users can interact with each other and with a range of social and traditional media platforms in realistic ways, these interactions are not measured or recorded in such a way as to allow the collection of empirical evidence for the effects of specific messaging strategies or operational tools.

The Team N's pilot simulation (currently named "Discovery") provides a single platform that can be modified to increase similarity to various social media platforms, facilitates interaction between users, and collects data on attention to and interaction with specific messages. It is designed to be expandable and adaptable so that it can ultimately link with a range of datasets and other CSPG components.

In addition, COGSEC is developing a pilot research study to explore the kind of experimentation that can be carried out on this kind of simulation platform, and the challenges likely to arise in implementing applied cognitive security research. This study takes real-world social media datasets collected in previous ARLIS research in the wild, and adapts them to test hypotheses about the effect of emotional content on message sharing in a controlled environment. This exercises potential capabilities in integration of in-the-wild and in-the-lab methods, research design and approval using complex messaging, target population recruitment, engagement of users in an open-ended research task, and data collection and analysis. It addresses potential challenges to balancing the variability and complexity of realistic live simulation with the control needed to collect meaningful and interpretable data.

3.3 The Pilot Simulation Platform of "Discovery"

There are many social media simulators available for use in training. However, despite the high fidelity of some of these platforms, they lack key features to facilitate applied research in the same settings. Most notably, their recording and instrumentation capabilities are at the level required to provide feedback to trainees, and lack the detail required to collect rigorous data on how research populations interact with messages based on controlled characteristics of those messages. Conversely, existing platforms designed for research on social media responses are low-fidelity, facilitate minimal or no interaction among participants, or are difficult to access for applied intelligence and security research.

As one of the initial foundations of the CSPG, ARLIS is developing a new in-house platform specifically designed to execute this type of experiment, and capable of collecting detailed data on key measures of social media engagement. This highly customizable experimental social media platform - currently named Discovery - includes additional, comprehensive instrumentation of participant interaction with individual posts and news items to allow for more sophisticated analyses. Having this platform in-house provides additional advantages, principally the ability to rapidly tailor experimental design and data collection to specific customer goals. Built on a MySQL database and a web framework written in Python Flask, this version of the social media platform includes the following features:

- multi-participant interaction (sharing, liking, and commenting on media posts);
- the ability to instrument and track participant behavior (view time, clicks, mouse movements);
- extensibility (the ability to add, subtract, or modify components of the platform);
- high-fidelity front-end User Interface (UI); and
- the ability to ingest arbitrary social media datasets (including images and videos).

Discovery incorporates the Infrastructure as Code (IaC) approach and will presently utilize containerization services (e.g., Docker, Amazon ECS) in order to ensure scalability and reliability of the platform. The modularity of Discovery will facilitate the development of algorithms (e.g., NLP, bot) for large-scale simulations with consonant reproducibility. Consistent testing, staging, and production environments with the same configuration will allow the platform to host any foreseeable number of participants on the requisite number of experimental nodes.

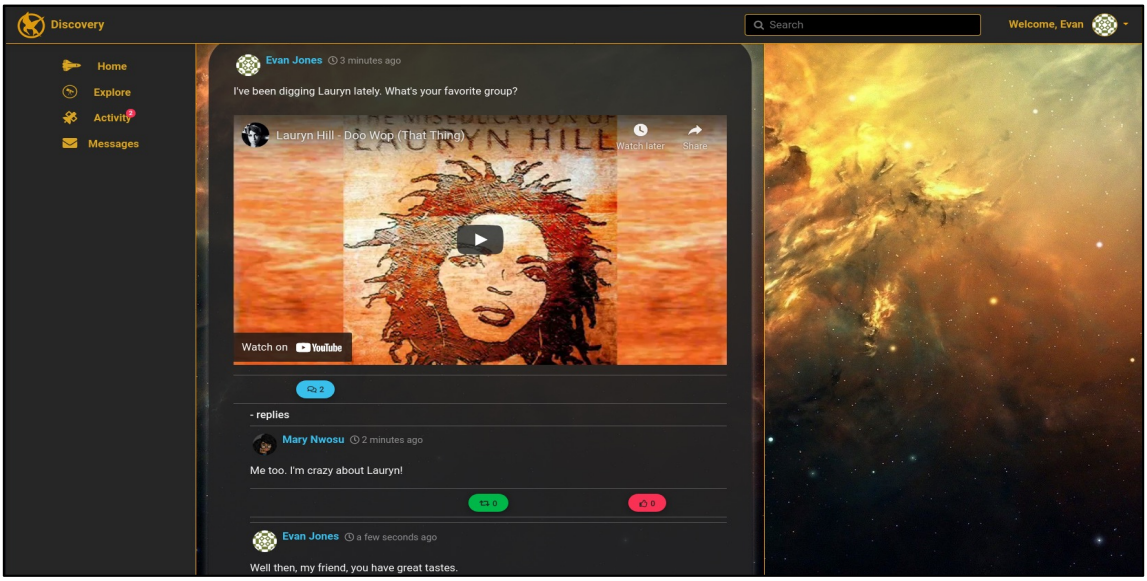

Fig. 3. A Snapshot of the Discovery Interface

4 THE SIMULATION MODEL: DISCOVERY

4.1 About Discovery

Team N's Cognitive Security Social Media Simulator, currently named "Discovery", is an experimental social media simulation platform capable of testing and evaluating multi-participant interaction in a high-fidelity environment as part of ARLIS's Cognitive Security Proving Ground (CSPG). The application is written primarily in Python Flask and is built on top of Miguel Grinberg's Microblog, a Flask app created by Miguel as part of his Flask Megatutorial.

The core functionalities of the simulator are [1] the ability to arbitrarily ingest real-world data sets via a REST API, [2] the ability to instrument and track user behavior in real time via Apache Flagon's UserALE, [3] the ability for users to message each other; post, comment, share like content (MySQL) [4] an administration ability to monitor, delete, and block content as necessary, [5] extensibility (the application relies on docker and docker-compose for deployment), and [6] a high fidelity front-end user interface (Material Bootstrap for templating).

The design principles aimed for were extensibility – how easily can we add, subtract, modify, and swap components of the app? Scalability – it works well with a few users, but can it run with 50, 100, 1000, or tens of thousands of users? And finally, reliability/reproducibility – it works on developer A and B's local environments, but will it work in production? If we decide to deploy it on bare-metal servers, can we? Can we deploy it easily on different cloud platforms (Azure, AWS, GCP)? Can we mix-and-match cloud resources if necessary?

These principles are closely related; part and parcel of a DevOps approach. With that in-mind, we followed an Infrastructure as Code (IaC) approach, utilizing Docker and containerizing as much of the app's components as possible. The idea is that this automatically enables extensibility, gets us 95 percent of the way towards guaranteeing reliability/reproducibility, and puts whomever is in charge of deployment on solid ground for easy scalability. With the help of docker-swarm, Amazon ECS or kubernetes, the app should easily be able to scale up to the requisite number of nodes for smooth functioning with varying user loads.

4.2 App Architecture

In production, the app itself consists of 9 different microservices running in unison. The web app itself can be run locally, outside of a docker container for testing development. However, its functionalities are limited in that case. Each of the containers are defined in docker-compose.yml which also reads in hyperparameters such as access tokens, connection URIs, etc. from a local .env file.

We use Python Flask as the web app back-end and the front-end consists of standard javascript and html. We use Material Design Bootstrap's css templates. The front-end is instrumented with Apache Flagon's UserALE which tracks and logs all user behavior in the browser. UserALE can utilize any back-end to store the logs, however we opted to use the standard Elasticsearch, logstash, and Kibana (ELK) stack for which the Flagon team provides a docker-compose file. More information on configuring UserALE and the ELK stack can be found in the Configuring UserALE section.

Images are hosted on a remote Amazon S3 server and stored as URLs in the MySQL database. The connection to the S3 server is specified in the .env file and instantiated in init.py using the boto3 SDK. However, in theory, any S3 server can work so long as the S3 credentials are changed in .env and an appropriate Python SDK is swapped out for boto3.

User information and posts are stored in a separate elastisearch server. The search bar is connected to this server so that users can search for content and other users. We use Redis and Redis Queue for background tasks. Finally, we use MySQL for the relational database to store all the models. The models are created with the use of Flask-SQLAlchemy

and so are completely independent from the SQL flavor chosen. Although we chose MySQL, this container can be swapped out for any flavor of SQL that is compatible with SQLAlchemy.

Currently, we are running the app on a single node. In production, however, the UserALE elastic stack should be on it's own node, at minimum (better spread across 3 or more nodes). Because the app is completely containerized it can be scaled up to any number of nodes necessary to handle traffic of any volume. We have left the choice up to users as to how they want to scale up and deploy the application (AWS ECS, docker-swarm, Kubernetes, etc.). In the future, we plan to include example scripts for how to deploy with these services.

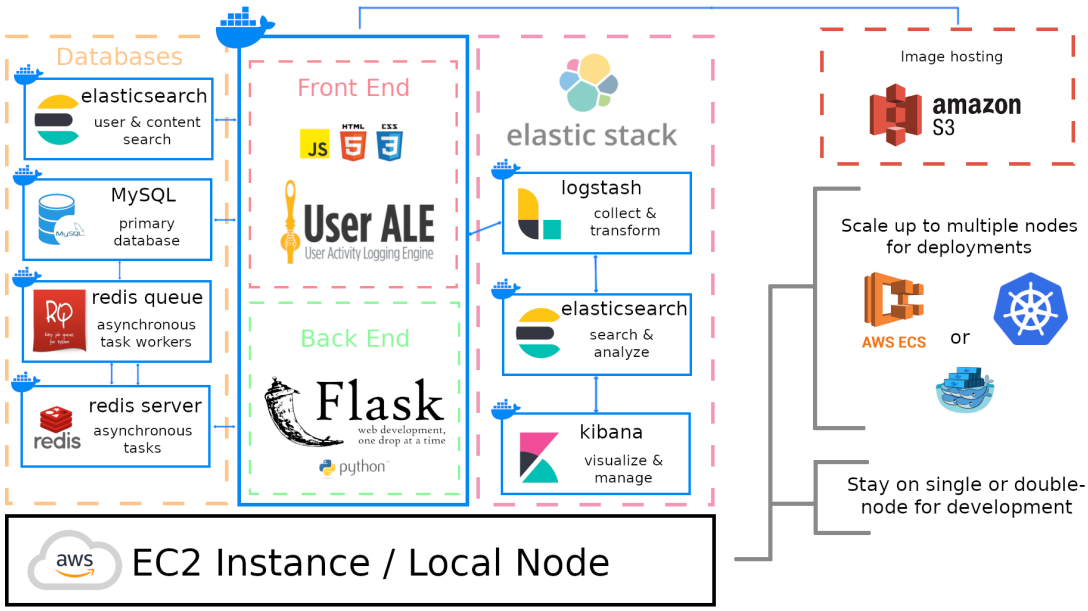

Fig. 4. The Application Architecture of Discovery

4.3 Database Schema

Figure 5 is a diagram of the MySQL relational database that Team N developed:

5 THE PILOT STUDY

5.1 Methods

Participants are recruited through an agency, which offers sign-up with its service opportunities to fill out surveys or complete studies for payment, letting them know beforehand only the length of the study in question. The aim was to bring in 100 participants, each from the general adult populations of Lithuania and Poland based on the power analysis described below. Interested participants connect to the social media simulation platform. Participants complete a consent form, then receive the following instructions translated into the relevant language:

"On the following site, you will be assigned a random account name and see social media posts from winter and spring 2021. Please read – and if you wish react to, respond to, or share – posts just as you would on an actual social

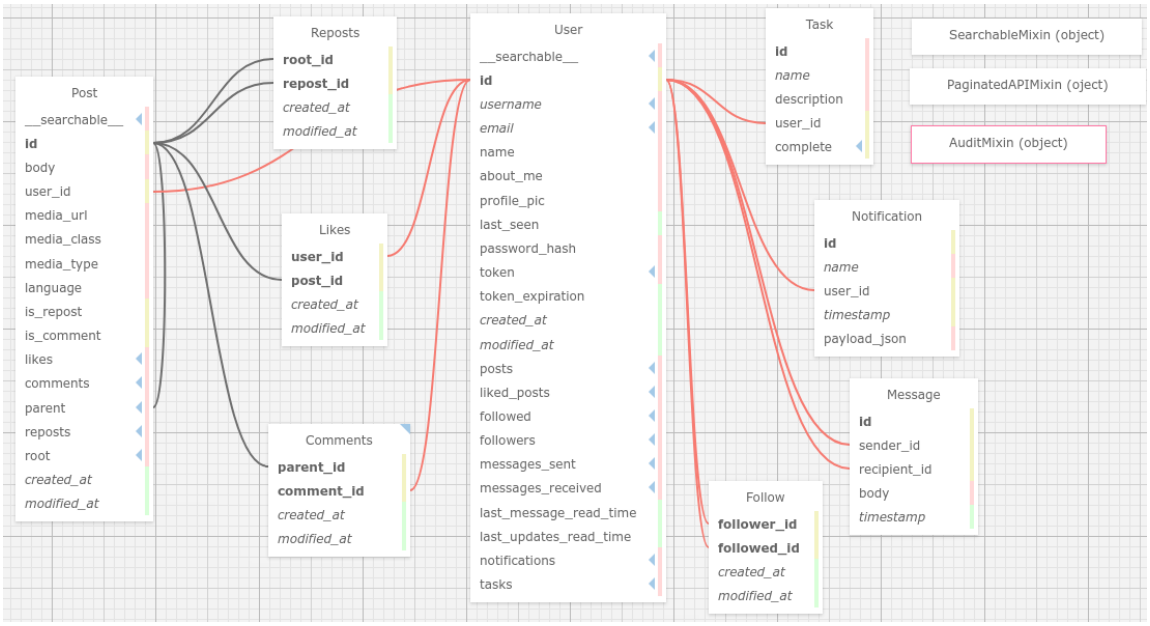

Fig. 5. The Database Schema of Disovery

media site. This experimental site is not connected to any real social media site, but other participants in the experiment will be able to see your likes, responses, and shares under your randomly assigned account name. At the end of 20 minutes you will be sent back to your survey site to receive credit and payment."

Following the instructions, participants will be presented with a simulated social media feed containing the 30 messages described above. Messages appeared approximately as they would in the original Facebook posts, with some visual differences due to the constraints of the simulation platform (e.g., images appear at small size and require a click to view at full size, original user icons are replaced with new ones assigned by the platform). Participants were able to browse at their own pace through available messages and choose to “share”, “like”, and reply to as many of them as they wished, with an experience similar to actual social media engagement. At the end of 20 minutes they will be redirected to a debriefing message, including links for in-country fact-checking sites related to the pandemic and vaccination options, and returned to the recruiting agency to receive their payment.

5.2 Analytic Plans and Implementation

We note that pilot studies are always limited in terms of the types of conclusions possible. This is a proof of concept pilot, and the results will be interpreted with caution. While we have planned for an appropriate number of participants for adequate power to interpret results, the materials themselves are limited. We note two main caveats to the interpretability of the results: [1] Given the time limits required for recruitment, participants will only be exposed to five posts labeled for each emotion, and [2] The social media post stimuli differ across the Polish and Lithuania samples and will not be tested or evaluated for equivalence across the two languages. This latter issue particularly limits interpretation of cross-country comparisons. In addition, time-limited nature of the pilot does not allow us to follow up with participants

for feedback on how their engagement motivations may differ from their behavior on real social media platforms. This type of qualitative input to the user experience will be a necessary component of future design development.

G*Power power analysis concluded that for a repeated measures between-subjects MANOVA, we would need a sample size of between 78 and 114 (per country) to achieve an acceptable power level between .80 and .95 for the following research questions.

5.2.1 *Research Question 1.* : For each emotion, how often are posts shared, liked, and replied to?

We calculated descriptive statistics for each of the three behavioral outcomes for the six emotions included in the experiment for both the Polish and the Lithuania samples. Participants either completed the post interaction behavior (sharing, liking, or replying) which was recorded as an interaction (coded as 2), or they will not complete the behavior which will be recorded as a non-interaction (coded as 1). The rate of behavior (for each: share, like, reply) was calculated as a percentage for each emotion and each country sample by dividing the number of interactions by the number of total possible interactions for each participant. The percentages were then averaged across all participants in the sample for the final rates. These descriptive statistics give a summary of how participants respond to various emotions overall, indicate general trends in the results, and comparatively how two cultures differentially respond to these same emotions.

5.2.2 *Research Question 2.* : What are the effects of each emotion (anger, pride, happiness, excitement, surprise, sadness) on sharing behavior (share, like, reply)?

To answer Research Question 2, we ran two repeated measures MANOVAs to examine the three post interaction behaviors (share, like, reply) as dependent variables (DVs), comparing across emotions (anger, pride, happiness, excitement, surprise, sadness) as the independent variable (IV). We calculated post-hoc comparisons across the 6 different emotions to compare interaction behaviors. We used the above calculated rate of behavior scores for each participant for the DVs. We ran separate analyses for the Polish and Lithuanian samples, so no cross-country comparisons will be possible here. This analysis allows us to compare rates of interaction behaviors with the social media posts (share, like, reply) across different emotional content of the posts.

5.2.3 *Research Question 3.* : Do patterns of sharing based on emotion differ between the Polish and Lithuanian samples?

To answer Research Questions 3, we ran a repeated measures within-between subjects MANOVA to examine the three post interaction behaviors (share, like, reply) as dependent variables (DVs), with emotion (anger, pride, happiness, excitement, surprise, sadness) and country (Poland, Lithuania) as independent variables (IVs). We calculated post-hoc comparisons across the 6 different emotions and two countries to compare interaction behaviors. We used the above calculated rate of behavior scores for each participant for the DVs. This analysis provides some preliminary insight into the intersecting effects of emotion and country - that is, do the effect of different emotions on interaction behaviors differ across different countries. This analysis is exploratory, and we note serious caveats to the interpretability of the results given that the social media post stimuli differ across the Polish and Lithuania samples, and the restricted number of stimuli per emotion participants were exposed to given the time constraints.

6 PERFORMANCE EVALUATION: BUILDING ON PILOT RESEARCH

6.1 Lessons Learned

Pilot studies are designed not only to collect informative data in their own right, but to identify potential issues and opportunities for larger-scale projects that build on them. This project, in particular, prototypes new methods as well as a new use of a platform for research, and new sources of recruitment for participants.

The primary lessons learned, in that context, are about timetable and data access. This pilot study was originally planned to leverage a preexisting dataset, and use that to quickly begin the human subjects approval process. However, the dataset turned out to come with its own constraints and limitations which had to be resolved prior to finalizing our choice of posts to be used as stimuli, which slowed the IRB application process. In particular, describing the size and characteristics of the dataset, and determining which subset met our criteria for inclusion, required considerable discussion and negotiation with the original research team.

CSPG studies are likely to depend on this type of preexisting social media dataset to place research participants in realistic simulations. In some cases, these datasets will be usable with minimal subset collection and cleaning (e.g., with the goal of measuring responses to specific keywords or hashtags easily searchable among the posts). In many, however, we will require human annotation and/or automated labeling prior to study implementation (e.g., to measure response to extra-textual characteristics of messages, such as suspected association with an influence campaign). These are time-consuming processes, so existing datasets that have already been labeled will be extremely valuable. Our experience here demonstrates, however, that even pre-labeled datasets will likely require additional work to adapt for CSPG purposes, and this should be taken into account in planning timelines. Studies may often require small-scale pilots prior to full collection in order to ensure that messages and designs are meeting requirements.

Another aspect of the study that increased the timeline, one likely to be common in future research, is the use of non-U.S. populations and the need to translate materials to and from relevant languages. We were fortunate to have access to the Lithuanian and Polish language experts from the original Minerva project, which minimized translation time. When external translators are being hired, care should be taken to retain them throughout the experimental preparation and implementation process. We found that we regularly needed to reach out for "just one more thing" as we added labels to the simulated social media platform (e.g., reply button), clarified instructions (e.g., making clear whether participants would receive credit when clicking an "Exit" button), or were asked for changes to facilitate the connection with the recruitment platform (e.g., referring to the company name used locally). HRPO approval also requires a signed statement of translation accuracy from the translators, for which we now have the template.

Recruiting in relevant target populations can be challenging, even relatively accessible samples (e.g., with internet access and interested in research participation) are likely to require local expertise to connect with. Earlier work at ARLIS attempted to use recruiting services such as in-house services, and found that they did not perform as well even with U.S. non-English samples as with U.S. English samples (e.g., a small study comparing English-speaking and Spanish-speaking populations was unable to bring in sufficient bilingual or primarily-Spanish participants to complete an analysis). Once contracted and connected with our research platform, the current recruiters were able to bring in participants at a rapid pace, getting us to 100 participants in Poland over a single weekend and in Lithuania. However, the 20-minute limit on their studies is likely to be a barrier to many important types of CSPG research, including those that combine simulation platform engagement with other forms of data collection, and those that seek to study multi-person engagement over extended periods of time. ARLIS is building direct relationships with institutions and collaborators in several nations, but in many cases will still need to identify and subcontract to a range of recruitment

services with local expertise. Creating and vetting a directory of such providers would be a significant benefit to future CSPG research.

Finally, one key positive lesson has been the importance of interdisciplinary expertise to support our technical methods. Our team and the study have benefited from the inclusion of and consultation with experimental psychologists, wargame designers, data scientists, computer scientists, and linguists. This has been valuable not only because each of these disciplines has specific contributions to make to the final study, but in the multiple perspectives brought to overcoming barriers and solving problems, and in our ability to question each others' discipline-specific assumptions as we sought to create a stronger product. Some of our most useful insights have started with, "This is probably a dumb question, but…" We anticipate that future CSPG research will continue to draw on this kind of interdisciplinary strength.

6.2 Discussion

The work carried out in this current pilot informs further development of CSPG infrastructure as well as studies supporting customer needs. In the short term, both these lines of work are directly continuing in the context of COGSEC. One of the goals of CSPG was to better detect and characterize malicious messaging campaigns online, including improved characterization of the populations that engage with and are persuaded by such campaigns. Team N is one of the teams drawing from academia and industry developing tools to meet these goals.

Previous methods for this type of evaluation have compared the results from novel tools to baseline annotated datasets (i.e., with best-current-practice ground truth labeled by humans and/or existing automated techniques). However, this creates a paradox: when the results of novel tools differ from TE results, there is no uncontroversial way of determining whether those tools have failed to meet the standard of current practice, or succeeded in outperforming it. The simulation-based methods explored here provide an alternative. ARLIS is supplementing standard annotation techniques by:

- Using subsets of social media data with known influence campaign content in experiments similar to those described here. Participants from target populations will fill out demographic and psychographic questionnaires, then interact with posts on the Discovery platform. We will measure their engagement with influence campaign content and the relationship between that engagement and questionnaire responses, allowing direct access to ground truth about responding populations; and
- Inserting TE-created "red team" campaigns into larger social media datasets on the simulation platform in a wargame-like activity, outputting a hybrid dataset that includes both real and synthetic posts. Analysis of this dataset using the novel tools can then be compared with direct red-team records of influence campaign content.

We are also working to make these methods available for a wide range of ARLIS projects and applied research goals, as well as to integrate the Discovery platform with other CSPG components (e.g., for modeling and simulating the effect of social interactions and adversary efforts beyond social media) for more complex research as well as wargaming use cases.

7 CONCLUSION

The objective of this pilot, as discussed, is to illustrate the potential for sophisticated, groundbreaking research in the cognitive security realm using a multi-method approach that incorporates longitudinal, empirical examination of real world data and controlled experiments with human subjects. The designed experimental study offers a glimpse

of what can be achieved with increased scale and methodological sophistication. Future work will build on lessons from the current pilot in order to scale up the number and variety of posts included in the simulation setting, and where necessary the number of participants involved. With the current foundation these advances appear feasible in the short term, with live experimentation and in-the-wild analytic components running somewhat ahead of modeling components.

[2–4, 6, 8, 10, 12–15, 18, 20, 25–30, 32, 35–38]

REFERENCES

[1] Mabrook S. Al-Rakhami and Atif Al-Amri. 2020. Lies Kill, Facts Save: Detecting COVID-19 Misinformation in Twitter. *IEEE Access* 8 (August 2020), 155961–155970. https://doi.org/10.1109/ACCESS.2020.3019600

[2] Christopher Andrew Bail. 2016. Combining Natural Language Processing and Network Analysis to Examine How Advocacy Organizations Stimulate Conversation on Social Media. *PNAS* 113 (Oct. 2016), 11823–11828. Issue 42. https://doi.org/10.1073

[3] Jason Brownlee. 2020. *Deep Convolutional Neural Network for Sentiment Analysis (Text Analysis).* Retrieved November 26, 2021 from https://machinelearningmastery.com/develop-word-embedding-model-predicting-movie-review-sentiment

[4] Jason Brownlee. 2021. *How to Use Word Embedding Layers for Deep Learning with Keras.* Retrieved November 7, 2021 from https://machinelearningmastery.com/use-word-embedding-layers-deep-learning-keras/

[5] Jyoti Choudrie, Snehasish Banerjee, Ketan Kotecha, Rahee Walambe, Hema Karende, and Juhi Ameta. 2021. Machine learning techniques and older adults processing of online information and misinformation: A covid-19 study. *Computers in Human Behavior* 119 (Jan. 2021), 11 pages. https://doi.org/10.1016/j.chb.2021.106716

[6] Kaggle: InClass Prediction Competition. 2018. *Fake News: Build a system to identify unreliable news articles – Dataset.* Retrieved November 7, 2021 from https://machinelearningmastery.com/use-word-embedding-layers-deep-learning-keras/

[7] Kate Cox, Linda Slapakova, and William Marcellino. 2020. *A Machine Learning Approach Could Help Counter Disinformation.* Retrieved October 11, 2021 from https://www.rand.org/blog/2020/06/a-machine-learning-approach-could-help-counter-disinformation.html

[8] Google Developers. 2021. *Machine Learning Glossary.* Retrieved November 7, 2021 from https://developers.google.com/machine-learning/glossary#recurrent-neural-network

[9] Cambridge Dictionary. 2021. *Fake News.* Retrieved November 20, 2021 from https://dictionary.cambridge.org/dictionary/english/fake-news

[10] Runa Ganguli, Akash Mehta, and Soumya Sen. 2020. A Survey on Machine Learning Methodologies in Social Network Analysis. In *2020 8th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO 2020).* IEEE, New York, NY, 485–489. Noida, India.

[11] Amira Ghenai and Yelena Mejova. 2017. Catching Zika Fever: Application of Crowdsourcing and Machine Learning for Tracking Health Misinformation on Twitter. In *2017 IEEE International Conference on Healthcare Informatics (ICHI).* IEEE, New York, NY, Article arXiv:1707.03778v1, 11 pages. https://doi.org/10.1109/ICHI.2017.58.

[12] Gaurav Goel. 2020. *Why Data is represented as a 'Vector' in Data Science Problems: Applications of Vector Algebra in Data Science.* Retrieved November 26, 2021 from https://towardsdatascience.com/why-data-is-represented-as-a-vector-in-data-science-problems

[13] Mohammed Zoher Guniem. 2021. *Identifying and Classifying Social Influencers – Engagement Analysis.* Master's thesis. University of Stavanger, Stavanger, Norway.

[14] Yuning Guo, Jianxiang Cao, and Weiguo Lin. 2020. Social Network Influence Analysis. In *2019 6th International Conference on Dependable Systems and Their Applications (DSA 2019).* IEEE, New York, NY, 517–518. https://doi.org/10.1109/DSA.2019.00093 Harbin, China.

[15] Saurabh Gupta. 2020. *Streamlit Web API for NLP: Tweet Sentiment Analysis.* Retrieved November 26, 2021 from https://www.analyticsvidhya.com/blog/2020/12/streamlit-web-api-for-nlp-tweet-sentiment-analysis/

[16] Ruth Harris, William Marcellino, and Linda Slapakov. 2020. *Using machine learning to detect malign information efforts online.* Retrieved October 11, 2021 from https://www.rand.org/randeurope/research/projects/using-machine-learning-to-detect-malign-information-efforts.html

[17] Michael Robert Haupt, Jiawei Li, and Tim K. Mackey. 2021. Identifying and characterizing scientific authority-related misinformation discourse about hydroxychloroquine on twitter using unsupervised machine learning. *Big Data Society* 1-15 (2021), 11 pages. https://doi.org/10.1177/20539517211013843

[18] Vladislav Karyukin, Aidana Zhumabekova, and Sandugash Yessenzhanova. 2020. Machine Learning and Neural Network Methodologies of Analyzing Social Media. In *ICEMIS '20 (DTESI '20).* ACM, New York, NY, 6 pages. https://doi.org/10.1145/3410352.3410739 Almaty, Kazakhstan.

[19] Katarina Kertysova. 2018. Artificial Intelligence and Disinformation. *Security and Human Rights* 29 (2018), 55–81. https://doi.org/10.1163/18750230-02901005

[20] Kajal Kumari. 2021. *Detecting Fake News with Natural Language Processing.* Retrieved November 5, 2021 from https://www.analyticsvidhya.com/blog/2021/07/detecting-fake-news-with-natural-language-processing/

[21] Tim K. Mackey, Vidya Purushothaman, Michael Haupt, Matthew C. Nali, and Jiawei Li. 2021. Application of unsupervised machine learning to identify and characterise hydroxychloroquine misinformation on Twitter. , e72–e75 pages. Commentary.

[22] Rivali Mamidi, Michele Miller, Tanvi Banerjee, William Romine, and Amit Sheth. 2019. Identifying Key Topics Bearing Negative Sentiment on Twitter: Insights Concerning the 2015-2016 Zika Epidemic. *JMIR Public Health AND Surveillance* 5, 2 (2019), 18 pages. https://doi.org/10.2196/11036

[23] William Marcellino, Kate Cox, Linda Slapakova, Amber Jaycocks, and Ruth Harris. 2020. *Human-machine detection of online-based malign information.* RAND Corporation Research Report RR-A519-1. RAND Corporation, Santa Monica, CA.

[24] William Marcellino, Christian Johnson, Marek N. Posard, and Todd C. Helmus. 2020. *Foreign Interference in the 2020 Election: Tools for Detecting Online Election Interference.* RAND Corporation Research Report RR-A704-2. RAND Corporation, Santa Monica, CA.

[25] Henrik Lesso Mjaaland. 2020. *Detecting Fake News and Rumors in Twitter Using Deep Neural Networks.* Master's thesis. University of Stavanger, Stavanger, Norway.

[26] NVIDIA. 2021. *Deep Learning Institute: Fundamentals of Deep Learning Course.* Retrieved October 10, 2021 from https://courses.nvidia.com/courses/course-v1:DLI+C-FX-01+V3/courseware/

[27] William Ogallo and Andrew Kanter. 2016. Using Natural Language Processing and Network Analysis to Develop a Conceptual Framework for Medication Therapy Management Research. *AMIA Annual Symposium Proceedings 2016* 2016 (2016), 984–993.

[28] Christopher Olah. 2015. *Understanding LSTM Networks.* Retrieved November 7, 2021 from https://colah.github.io/posts/2015-08-Understanding-LSTMs/

[29] Ray Oshikawa, Jing Qian, and William Yang Wang. 2020. A Survey on Natural Language Processing for Fake News Detection. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020).* European Language Resources Association, Luxembourg, 6086–6093.

[30] Lakshmi Panneerselvam. 2021. *Build Your Own Fake News Classifier With NLP.* Retrieved November 26, 2021 from https://www.analyticsvidhya.com/blog/2021/06/build-your-own-fake-news-classifier-with-nlp

[31] Alejandro Rodríguez-González, Juan Manuel Tuñas, Lucia Prieto Santamaría, Diego Fernández Peces-Barba, Ernestina Menasalvas Ruiz, Almudena Jaramillo, Manuel Cotarelo, Antonio J. Conejo Fernández, Amalia Arce, and Angel Gil. 2020. Identifying Polarity in Tweets from an Imbalanced Dataset about Diseases and Vaccines Using a Meta-Model Based on Machine Learning Techniques. *Applied Sciences* 2020, 10, Article 9019 (Dec. 2020), 13 pages.

[32] Ben Roshan. 2020. *Fake News Classifier on U.S. Election News: LSTM.* Retrieved November 26, 2021 from https://www.analyticsvidhya.com/blog/2020/12/fake-news-classifier-on-us-election-news

[33] Richard F. Sear, Nicolas Velasquez, Rhys Leahy, Nicholas Johnson Restrepo, Sara El Oud, Nicholas Gabriel, Yonatan Lupu, and Neil F. Johnson. 2020. Quantifying COVID-19 Content in the Online Health Opinion War Using Machine Learning. *IEEE Access* 8 (May 2020), 91886–91893. https://doi.org/10.1109/ACCESS.2020.2993967

[34] Linda Slapakova and William Marcellino. 2020. Expert Insights with RAND Europe: Using AI to Tackle Disinformation Online. Audio. Retrieved October 11, 2021 from https://www.rand.org/multimedia/audio/2020/10/23/using-ai-to-tackle-disinformation-online.html

[35] Qiaoyu Tan, Ninghao Liu, and Xia Hu. 2019. Deep Representation Learning for Social Network Analysis. *Frontiers in Big Data* 2 (April 2019), 10 pages. Issue 2. https://doi.org/10.3389

[36] Sai Teja. 2020. *Stop Words in NLP.* Retrieved November 24, 2021 from https://medium.com/@saitejaponugoti/stop-words-in-nlp-5b248dadad47/

[37] Soroush Vosoughi, Deb Roy, and Sinan Arul. 2018. The Spread of True and False News Online. *Science* 359 (March 2018), 6 pages.

[38] Kai-Cheng Yang, Onur Varol, Clayton A. Davis, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. 2018. Arming the Public with Artificial Intelligence to Counter Social Bots. *Human Behavior Emerging Technologies* 2019 (Dec. 2018), 48–61. Issue 1. https://doi.org/10.1002/hbe2.115

[39] Yuehua Zhao, Jingwei Da, and Jiaqi Yan. 2020. Detecting health misinformation in online health communities: Incorporating behavioral features into machine learning based approaches. *Information Processing and Management* 58, 2020 (Sept. 2020), 24 pages. https://doi.org/10.1016/j.ipm.2020.102390