Article

SM+S
social media + society

Social Media + Society
October-December 2023: 1–22

DOI: 10.1177/20563051231216964
journals.sagepub.com/home/sms
Sage

Twitter Propaganda Operations: Analyzing Sociopolitical Issues in Saudi Arabia

Craig Douglas Albert, Ahmed Aleroud, Yufan Yang, Abdullah Melhem, and Josh Rutland

Abstract

The purpose of this article is to explore Arabic-language Tweets based out of Saudi Arabia to investigate the social media landscape. Specifically, this article seeks to address the question, "What thematic issues concerning the U.S. sociopolitical landscape are present in Arabic-language Twitter postings?" And, "To what extent can these issues be described as propagandic in nature?" To do so, we propose a machine-learning and artificial intelligence span detection approach to identify propaganda Tweets in Middle Eastern Countries, with a focus on Saudi Arabia. As opposed to previous work, this article maps and investigates different propaganda categories using the BEND Social Cyber Security framework. This article then proceeds to a case study analysis of state-sponsored targeted propaganda from Saudi Arabia and briefly describes the categories of propaganda uncovered. We then relate those categories to the BEND Framework and conclude with policy recommendations and discussion.

Keywords

machine learning, artificial intelligence, social media analysis, BEND framework, the Middle East, Saudi Arabia

Introduction

Social media platforms have become a battlefield for propagandic campaigns both within and across state borders. Research has shown that both party elites and foreign rival countries have been using social media to spread misinformation, disinformation, and propaganda to serve their goals, such as stigmatizing political opponents and interfering with a rival country's domestic politics (Badawy et al., 2018; Benkler et al., 2018). However, studies on propaganda campaigns between rival states largely focus on the United States–Russia and the United States–China, with only a few researchers exploring Middle Eastern countries' propaganda efforts despite the fact that anti-American sentiment widely exists in Middle Eastern countries (Jamal et al., 2015). Nevertheless, anti-American sentiment plays a critical role in great-power politics such as U.S.–Russia (Mendelson & Gerber, 2008) and U.S.–China relations (Weiss, 2013) and thereafter global peace, as well as in transnational terrorist attacks targeted at the United States (Neumayer & Plümper, 2011). Although there is a long way between increasing hostility toward a foreign country to an open war between nation states, research has shown that increasing hostility among rivalries is associated with more terrorist attacks from one to another (Conrad, 2011). Furthermore, various U.S. documents suggest that foreign propaganda threatens the state by undermining national security objectives of the United States and its allies, jeopardizing trust in democracy, fueling political unrest and violence, and destabilizing society (Chernobrov & Briant, 2022). Although there has been a long-standing tension between some Middle Eastern countries and the United States, Arabic influence operation and propaganda strategies targeted at the United States are understudied.

Exploring how Middle Eastern countries frame the United States to their citizens and what propagandic strategies are implemented in the framing process will deepen our understanding of elite behavior and the public sphere regarding the United States in Middle Eastern countries, which is

Augusta University, USA

Corresponding Author:
Craig Douglas Albert, Master of Arts in Intelligence and Security Studies, Augusta University, 2500 Walton Way, Augusta, GA 30904, USA.
Email: calbert@augusta.edu
X: @drcraigdalbert

beneficial to American national security. Our research attempts to provide more evidence on Middle Eastern countries' propaganda efforts and influence operations regarding the United States on social media. Specifically, we ask, "What thematic issues concerning the U.S. socio-political landscape are present in Arabic-language Twitter postings?" And, "To what extent can these issues be described as propagandic and state-sponsored influence operations in nature?" To do so, we analyze state-sponsored Twitter data from Middle Eastern countries, identifying the propaganda techniques with the help of machine learning algorithms, and explore the content of these Tweets using topic modeling. We also add a case study of Saudi Arabia to illustrate our findings. Before delving into the review of literature, it is necessary to conceptualize these terms.

We fit propaganda as a type of influence operation within the larger context of information warfare. In other words, information warfare is an operational objective, aimed at "the deliberate manipulation or use of information by one party on an adversary to influence the choices and decisions the adversary makes in order for military or strategic gain" (Whyte et al., 2020, p. 344). Within this operational environment exists several tactics or tools to achieve a goal. These are the influence operations, which when tied to cyberspace and social media can be referred to as cyber-enabled influence operations (CEIO). As Nakayama (2022) writes, "information operations that leverage means and dynamics unique to cyberspace—with a particular focus on operations targeting social media" (p. 50). Within this realm, propaganda is the spreading of messages through social media channels and is a type of CEIO. As Bastos and Farkas (2019) demonstrate, "Propaganda campaigns are often implemented by state actors with the expectation of causing or enhancing information warfare" (p. 3). Thus, there is a nexus between influence operations and propaganda. In other words, in studying propaganda, this study also fits within the larger context of influence operations and social media information warfare. Thus, to understanding propaganda, one must understand influence operations in general (Cordey, 2019).

There are several understandings of influence operations especially on social media. Perhaps the most significant is Larson et al. (2009), who argue that in the context of U.S. national security, influence operations, "are the coordinated, integrated, and synchronized application of national diplomatic, information, military, economic, and other capabilities in peacetime, crisis, conflict, and postconflict to foster attitudes, behaviors, or decisions by foreign target audiences that further U.S. interests and objectives" (p. 2). Larson et al. (2009) argue that these operations influence a target audience, whether an individual leader, members of a decision-making group, military organizations and personnel, specific population subgroups, or mass publics (p. 2). In general, however, influence operations refer to intelligence operations that interfere in the affairs of another actor (Callanan, 2009; Maschmeyer et al., 2023). Many researchers argue that what is important is that the decision-making capabilities are interfered with through influence operations and information warfare. As Theohary (2018) makes clear, "Whether attacking government agencies, political leadership, or news media to influence public opinion or to complete decisionmakers to take certain actions, ultimately the target of information warfare activities is human cognition" (p. 2). Building on this, Bergh (2020) defines influence operations as a concerted effort by an actor, such as a state or terrorist group, to interfere in the process and meaning making by individuals or groups outside its own legal control through tools and facilities on publicly available social media services (p. 111). An influence operation is, therefore, an umbrella term or, as Cohen and Bar'el (2017) have put it, "a catchall phrase for any action intended to galvanize a target audience—an individual, a prominent group, or a broad audience—to accept approaches and to adopt decisions that mesh with the interest of the instigators of the operation" (p. 13).

As an umbrella term, influence operations include all types of operations in the information domain, not only propaganda operations but also clandestine and intrusive activities such as cyber-espionage and cyberattacks (Brangetto & Veenendaal, 2016). Among all these influence operations tactics, we focus on propaganda in this research because of the following reasons. First, focusing on propaganda allows us to speak to broader audiences and contribute to the rich literature of war and propaganda dating back to Harold D. Laswell, Edward Bernays, and Walter Lippmann. Second, compared with the umbrella terms, such as influence operations and information operations, the concept of propaganda is more operationalizable because studies from both political science and discourse analysis have provided various theoretical frameworks and analytical tools (for example, see Jowett and O'Donnell's, 2018, conceptualization and typology of propaganda and Van Dijk's, 2011, discussion of logical fallacies used in discourse analysis). Finally, compared with other influence operation tactics such as cyberattacks, data on propaganda operations are more available because of social media. Therefore, we focus on state-sponsored propaganda operations on social media, which consists of coordinated accounts exploiting the online space to influence public opinion (Ng et al., 2022) to serve the interest of a state.

Regarding propaganda, we borrow Jowett and O'Donnell's (2018) definition that propaganda is "the deliberate, systematic attempt to shape perceptions, manipulate cognitions, and direct behavior to achieve a response that furthers the desired intent" (p. 2). This definition emphasizes that (1) propaganda is deliberate, meaning that it serves a certain purpose favoring the propagandist's interest; (2) propaganda is systematic, indicating that the propaganda efforts are usually large-scale and persist over time; and (3) the goal of propaganda is to manipulate the audience. Therefore, when a state implements

purposeful propagation of an idea or narrative intended to influence a target audience (Theohary, 2018) to shape the audience's perception and manipulate public opinion, it can be understood as a form of influence operations.

The remainder of this article proceeds in the following order. First, we provide a basic review of social media influence operations and propaganda, especially through state-sponsored activities. We then turn to our modeling and research design elements and describe how the technical aspects were implemented. Next, we present our case study analyses and provide case-specific results pertaining to the Kingdom of Saudi Arabia. Finally, based on these results, we provide policy recommendations and discuss holistically why understanding social cybersecurity in the context of Arabic tweets affects the policy realm of U.S. national Security.

Literature Review

Twitter is well known as a domain for the execution of information operations and the dissemination of propaganda to global audiences (Arif et al., 2018; Starbird et al., 2019; Uyheng et al., 2020). The platform's ability to both monitor and amplify information that "trends" or generates significant public interest makes it an ideal breeding ground for infiltrating and directing existing trends (Prier, 2017) or spreading one's own artificially generated trend to a wider audience (Guarino et al., 2020). These operations are often complex in nature and make use of cultivated networks of nodes known to be efficient at communicating a narrative to the largest possible audience in the shortest amount of time (Guarino et al., 2020). They may also be carried out by anyone with internet access, from individuals like Farah Baker (Wolf, 2015) to terrorist groups (Weimann, 2010) and state governments (Kießling et al., 2020). Understanding the success of these operations requires an explanation of the networks that make them possible and how they are employed. This section explores these networks, their creation, and the importance of central nodes or "influencers" and bots. An analysis of successful influence operations carried out on Twitter follows. These examples include terrorist efforts by organizations such as ISIS and al-Shabaab, campaigns carried out by states such as Iran and Russia, and notable independent individual or small group–led efforts such as those of Farah Baker in Palestine and the recent QAnon conspirators.

The networks that make information operations possible, while reaching an extensive and often diverse audience, are often centralized around a specific and small group of individuals who are considered an ideological and informational authority within their immediate community of like-minded followers (Dilley et al., 2022; Guarino et al., 2020). These individuals, or "influencers," use their communal authority to disseminate information to their followers, who share it with their own associates, with the eventual goal being to generate attention around articles of information (Guarino et al., 2020). If done correctly, this information may then generate enough attention that web users on twitter who are outside the typical spheres of influence will be exposed to the information disseminated (Prier, 2017). If those users find the information interesting and begin to interact with and spread that information, it creates what is commonly referred to as a "trend" (Prier, 2017), which will then be broadcast to a global audience by Twitter's front page.

Ideally, trends are intended to be determined by interest of Twitter's human users, but it is possible for these trends to be artificially constructed or "hijacked" as described by Prier (2017). The hijacking as described by Prier (2017) is carried out by directing swarms of bots to generate artificial clicks and abuse the algorithm employed by Twitter to seek out and promote popular tweets, articles, and messages. Prier's (2017) hijacking typically references the steering of existing narratives toward an interested party's perspective of events, but it should be noted that propaganda does not require an existing trend to spread.

Various influencers (Dilley et al., 2022) or "hubs" scattered (Caldarelli et al., 2020, p. 2) across Twitter's userbase command swarms of bots that promote their ideas to disseminate them to wider audiences. These "bot squads" are "groups of bots that follow and retweet the same group of hubs," thus amplifying their messages and making them more likely to trend (Caldarelli et al., 2020, p. 2). The nature of users to associate with "strongly clustered" communities "sharing similar ideas" to their own on Twitter helps boost the effectiveness of this strategy, essentially enabling hubs to generate propaganda trends that spread misinformation and disinformation virally (Caldarelli et al., 2020, p. 2). However, the impact of the message and its resulting influence is not limited to the "true believers" who affiliate themselves with ideologically aligned hubs (Prier, 2017, p. 58). Even those who exist in the "outside network" may be influenced by the propagandic messages, despite "not necessarily subscrib[ing] to the underlying beliefs that support the narrative" (Prier, 2017, p. 58). An infamous example of this phenomenon was seen as a part of the so-called "astroturfing" method used by QAnon conspiracy theorists such as Jason Sullivan, Ron Watkins, and Jim Watkins (Dilley et al., 2022, p. 8), all of whom served as "hubs" for the larger QAnon movement. Despite frequent association with Twitter propaganda and bot usage, it should be emphasized that "hubs" and bot usage are not limited to conspiracy theory style propaganda, any singular ideology, or even any specific actor type. In fact, despite the argument that "such platforms [as Twitter] democratize public discourse, recent years have shown how adversarial actors may employ diverse strategies to manipulate public opinion toward disruptive social and political outcomes" (Uyheng et al., 2020). Independent individuals, states such as Russia, Iran, North Korea, and China (Ferrara, 2020, p. 11) as well as terrorist organizations such as ISIS (Moriarty, 2015) have been known to employ this tactic. The following explore a few examples involving each of these actor types.

As the Westgate mall terror attack of 2013 was occurring, the terrorist organization responsible, al-Shabaab, released over 500 tweets (Mair, 2016). These tweets claimed responsibility and shared continuous updates and live messages to keep the public's focus on them (Mair, 2016). Throughout the 4-day siege, Twitter served as the primary method of communication between the attackers, the government, first responders, and the Kenyan public (Simon et al., 2014). Sullivan's rhetorical analysis of the tweets released during the attack describes the terrorist group's behavior as performative, with the intent clearly being to persuade the local populace to hear their message and support their efforts (Sullivan, 2014). Ultimately, Mair concludes that al-Shabaab's goal was two-fold: maintaining public interest in the attack and controlling the narrative, though their efforts were geographically targeted (Mair, 2016). However, they are not the only group to weaponize twitter to boost their message.

The advent of the Islamic State's (ISIS) rise to power and the coinciding "cyber jihad" aptly demonstrate the effectiveness of Twitter propaganda (Singer & Brooking, 2018). Horror stories emerged across the world of radicalized youth pledging allegiance to ISIS's cyber operations arm, with many engaging in Twitter activism for the group by sharing its ideological messages, videos of beheadings, and fear mongering against the terrorist organization's adversaries (Mitts, 2019; Singer & Brooking, 2018). The radicalization efforts adopted by the terrorist group drew countless headlines and significant scholarly attention, particularly as the group found success among Western audiences that would ordinarily have been unlikely targets and associates of a geographically and culturally foreign movement (Mitts, 2019). The leap from geographically localized terrorist Twitter propaganda such as that seen in the Westgate attack (Mair, 2016) to the larger, globally oriented propaganda efforts seen with ISIS (Singer & Brooking, 2018) demonstrated the potential of Twitter propaganda. In the battle for narrative control, terrorist actors seemed to have evened the playing field for the states they opposed (Singer & Brooking, 2018). However, this is not to say that states have ignored the potential of Twitter and other social media operations.

Alizadeh et al. (2020) demonstrate that it is quite difficult to determine whether social media posts are organic or state-sponsored influence operations. However, the researchers developed a platform-agnostic supervised learning approach to classifying posts as being a part of a coordinated and thus state-sponsored, influence operation or not (Alizadeh et al., 2020, p. 9). They find that content-based features distinguish coordinated influence campaigns and were able to use their supervised learning approach to detect influence operations from Russia, China, and Venezuela across social media platforms. They found that Chinese operations were easier to notice than Russia, and Venezuelan were the easiest to determine state influence at work (Alizadeh et al., 2020, p. 3). In addition, Ng and Carley (2023) analyzed online conversations on Twitter about the Chinese balloon spotted in American airspace in January 2023 and identified that over 46.05% of the Chinese accounts involved in the conversations were bots, which is higher than the average proportion of bot population on social media. They also found that in these conversations, Chinese accounts focused on the shooting and removal of the balloon as well as using narratives that are related to former U.S. president Donald Trump, such as "MAGA" and "SleepyJoe."

Ng et al. analyze image-based influence operations from China, Iran, Russia, and Venezuela. Interestingly, they find three distinct lines of effort for Chinese operations and argue that "The Image-Image network's structure and high clustering coefficient may correlate to high-level coordination and integration of influence operations" (Ng et al., 2022, p. 6). For Iran, they find that Image-based tactics include suppressing political dissidents using political hate speech, vulgar speech, counter-speech, and religion and societal topics (p. 6). Russia's network was highly connected with evidence to drive division within the U.S. political landscape, NATO, and interesting, lifestyle themes of food and travel (p. 6). For Venezuela, their Image–Image network is more decentralized, and they focus on "Breaking News" to describe their own operations as news (p. 6). For all of them except China, there were images and memes of U.S. politicians, showing close correlation between Iran, Russia, and Venezuela with less Chinese connection (p. 6).

Focused on Russia's influence operations, Lukito (2020) argues that the Russian government–supported Internet Research Agency (IRA) produces and disseminates disinformation targeted at the United States across various social media platforms, including Facebook, Twitter, and Reddit, and IRA activities on Twitter are influenced by IRA activities on Reddit. The author further points out that it may be because the IRA is experimenting and trial ballooning on one Reddit to figure out the optimal information to distribute on Twitter. In addition, with the growing tensions between the United States and Russia, Chernobrov and Briant (2022) claim that these two countries have witnessed mutual accusations of disinformation and propaganda campaign targeted at each other, and the threat of disinformation campaign has become an important part in the relationships between the United States and its rivals.

Research on propaganda and information operations in the Arab world largely focused on conventional media and nonstate actors' use of social media. For example, Fahmy et al. (2012) have analyzed how satellite TV may shape public opinion in the Arab world, and Ali and Fahmy (2013) have explored the use of social media by protesters in Iran, Egypt, and Libya during the Arab Spring. Concerning state-sponsored Arabic social media influence operations, there seems to be much less academic literature. There are some key sources to highlight, however. DiResta et al. (2021) analyze Middle Eastern influence operations across social media networks and find a breadth of tactics and narratives, devoted to multiple geopolitical objectives-versus, for instance,

Chinese or Russian operations, that tend to focus on singular objectives or targets (p. 99). They conclude that influence operations are gaining steam and becoming more important in the MENA region.

Russia's interference campaign in the 2016 U.S. election is perhaps the most frequently cited, but the state's efforts extend far beyond this. There is evidence to suggest the state's involvement in the Brexit debate in England (Llewellyn et al., 2018), the vaccine debate surrounding COVID-19 in the United States (Broniatowski et al., 2022), and a myriad of other hot button issues at the center of Western politics (Miller, 2019). Beginning in 2022, much of Russia's propagandic focus has shifted to the Ukraine war, although Pierri et al. (2023) interestingly note a distinct drop in "the prevalence of Russian propaganda following the invasion" while also being careful to emphasize that the presence of propaganda "is not negligible" (p. 8).

Beyond Russia, Iran has been observed using Twitter bots and propaganda accounts to attempt to influence the foreign and public policies involving Saudi Arabia, with the state actively promoting biased hashtags and retweeting identifiable propaganda sources (Kießling et al., 2020). Kießling et al. (2020) note that Iran's Twitter campaign activity tends to spike coinciding with major political events, although they also find that the majority of Iran's attempts at influencing foreign policy were unsuccessful. China has also been observed employing Twitter propaganda strategies, notably surrounding public discourse concerning territorial disputes in the South China Sea, in which most of the 19 "important China actors in the #SouthChinaSea conversation" were identified as representing "central-level state news media" (Nip & Sun, 2022, p. 61). While each of these campaigns have varied in effectiveness and scale, their very existence demonstrates the popularity of social media influence operations and Twitter propaganda and among state actors. However, even individuals can employ these methods to create powerful effects.

One notable example belongs to Farah Baker, the widely acclaimed "Anne-Frank of Palestine" (Wolf, 2015). Baker began trending on Twitter in 2014 during the Israeli attacks on Gaza as part of its campaign against Hamas (McNeill, 2019). Baker was only 16 at the time, but her tweets rapidly evolved into one of the most successful narrative campaigns seen on Twitter (Patrikarakos, 2017). The narrative campaign rapidly picked up traction for its perceived authenticity and emotional gravity, allowing Baker to amass a substantial following (Patrikarakos, 2017). This following transformed Baker almost overnight from a humble teenager into an influencer hub with a vast network of like-minded supporters who spread her message to a global audience (Patrikarakos, 2017). This campaign ended up being so influential that it forced the Israeli Defense Force to reevaluate its own public relations and narrative influence strategies (Patrikarakos, 2017). A similar phenomenon occurred more recently with the spread of vaccine-related conspiracy theories on Twitter. About 65% of these conspiracy theories were traced back to a "Disinformation Dozen" of individuals running multiple accounts across the platform (Bond, 2021, para. 8). This handful of individuals managed to build massive community networks that spread propaganda across multiple social networks, including Twitter, to generate millions of retweets and views, eventually reaching many users outside their usual spheres of influence (Nogara et al., 2022). This network was so influential that U.S. Congressional Representatives and multiple state attorney generals repeatedly urged social media platforms, including Twitter, to ban the accounts of the "Disinformation Dozen". While these platforms took efforts to do so, the hubs were resilient in their creation of multiple new accounts to replace their lost ones.

Even within the world of Western democracies, intentional use of bots on social media to serve certain political purposes has been increasing. For example, Howard and Kollanyi (2016) find that political bots have been strategically used in the United Kingdom's 2016 referendum on leaving the European Union. Another European country, Germany, has also witnessed the use of bots by both political and private actors on social media to manipulate the public sphere and thereafter the 2017 German Federal Election (Neudert, 2018). In the United States, Howard et al. (2018) and Howard et al. (2017) have demonstrated the influence of bots, mis/disinformation, junk news, algorithms, and automated political communication in general on both local and federal-level elections. These examples demonstrate the power and prolificity of Twitter propaganda. Effective propaganda can be generated by almost anyone in the world, with the only requirement being an internet connection (Patrikarakos, 2017). Social media in general, but Twitter in particular, are likely to remain an influential component of public discourse surrounding major political events for the foreseeable future. As such, they will also remain an avenue to dispense propaganda and shape the course of narrative battles between states, interest groups, individuals, and even adversaries to the global system like ISIS. Now that a brief review of the literature has been presented, let us turn to the specifics of the current project.

Research Design

Twitter Data

This article addresses a new category of propaganda attacks that are tied to state-linked accounts that spread anti-U.S. propaganda by taking advantage of specific geopolitical crises in the Middle East. We investigated the role of general language models and general training data to detect those forms of targeted propaganda. Our general propaganda data are selected from a public data set. The state-linked data are selected from Twitter Moderation Research Consortium (TMRC), through which Twitter shares large-scale data on

information operations to the public since 2018. TMRC has published data sets of state-linked information operations originating from various countries, including Iran, Russia, China, Saudi Arabia, and more. For the purpose of this research, we focus on state-linked Twitter accounts' activities originating from Middle Eastern countries, including Saudi Arabia and Egypt. The following section describes each of both data sets in detail.

Data Collection

We used two distinct data sets to study both general and targeted propaganda. Data Set 1, "General propaganda" is a preexisting labeled data set consisting of a diverse collection of tweets sourced from well-known news outlets in Arab countries, supplemented by international news sources. This data set consists of 3,200 tweets from each source, with an additional 100 random tweets per source for augmentation, resulting in a sample of labeled 930 tweets. Data Set 2, referred to as "Targeted propaganda," was selected from Twitter's publication of a publicly available archive covering state-backed information operations from 2018 to 2022. The present research narrowed its focus within this data set to tweets related to the United States and predominantly characterized as anti-US propaganda. The data selected from this data set were not prelabeled. The labeling process encompassed both binary and multilabel classification, conducted by native Arabic-speaking annotators. The following two subsections describe each data set in detail and the labeling process for Data Set 2.

Data Set 1: General Propaganda

This data set was collected from the top news sources in Middle Eastern states, which include the social media pages for news sources such as Al Arabiya, Sky News Arabia from UAE, Al Jazeera, and Al Sharq from Qatar (Alam et al., 2022). Five international sources were added to those sources, including Al-Hurra News, BBC Arabic, CNN Arabic, and France24. The most recent 3,200 tweets from each source were selected. Another 100 random tweets were also used to augment each source. Then, a sample of 930 tweets for annotation were selected. As this is a multilabel classification problem, a skewed distribution is noticed in this data set as shown in Table 1.

Data Set 2: Targeted Propaganda

Between 2018 and 2022, Twitter published a comprehensive, public archive of data related to state-backed information operations. Thirty-seven data sets have been shared as a part of this effort. The published data sets attributed platform manipulation campaigns originating from 17 countries, spanning more than 200 million Tweets and nine terabytes of media.

As we were interested in targeted propaganda, we only focused on the tweets related to the United States and are mainly anti-United States in nature. The selected accounts were generally tweeting propagandic themes such as

- Muslim Brotherhood in the Middle East
- Iranians' role in the Middle East
- War in Syria
- Sanctions on Qatar

We created a dictionary to select all tweets related to the United States and U.S. institutions to include the U.S. military, the U.S. Army, and U.S. Armed forces in general. We also included keywords related to U.S. leaders who were active during the timeframe of tweets such as Donald Trump and Jared Kushner. The number of state-based tweets per each data set is shown in Table 2. Accounts were suspected to be linked to Saudi Arabia. We followed a rigorous procedure to label the data. As these are targeted propaganda tweets and need to be labeled against the 17 propaganda categories, the labeling process was time-consuming. It started in October 2022 and ended in January 2023. The labeling process consists of binary class classification and multilabel classification. The labeling process is demonstrated in Figure 1.

In the first phase, two annotators who are native Arabic speakers spent some time understanding the propaganda techniques in Arabic. They were given examples on each technique from the first General Propaganda Data Set 1 and many other publicly available examples on the web. The two annotators then conducted a binary classification to classify tweets into no propaganda labels (0) and (1), which is potential propaganda. Potential propaganda tweets were then labeled in the second phase to identify and extract contiguous spans of text that correspond to at least one propaganda technique. Propaganda span detection focused on identifying specific segments or spans of text within each tweet that contained elements of propaganda. Given that a single tweet can encompass multiple types of propaganda as shown in Table 1, span detection techniques allow to highlight and isolate these propagandistic phrases or sentences. This approach enables a more granular analysis, enabling us to better understand the various propaganda tactics employed in a single tweet and help in the development of effective countermeasures against the spread of disinformation and biased content on social media platforms.

In total we labeled about 222 as propagandic tweets in the first and second phase of labeling. Only 28 nonpropagandic tweets were found in the labeled data, which were excluded in the second phase. The two annotators hold graduate degrees in information systems. For the second phase labeling, we consulted a third annotator who is a Jordanian domain expert in the fields of political science, press, and media.

The average disagreement in the labeling processes in the multiclass stage using the Kappa index was 11%, which

Table 1. General Propaganda Data.

Propaganda technique	Description	No. of tweets
Appeal to authority	Stating that a claim is true simply because a valid authority or expert on the issue said it was true, without any other supporting evidence offered	28
Appeal to fear/prejudice	Seeking to build support for an idea by instilling anxiety and/or panic in the population toward an alternative	55
Black-and-white fallacy/ dictatorship	Presenting two alternative options as the only possibilities, when in fact more possibilities exist	3
Causal oversimplification	Assuming a single cause or reason when there are multiple causes for an issue	5
Doubt	Questioning the credibility of someone or something	30
Exaggeration/minimization	Either representing something in an excessive manner: making things larger, better, worse or making something seem less important or smaller than it really is	54
Flag-waving	Playing on strong national feeling (or to any group, for example, race, gender, political preference) to justify or promote an action or idea	7
Glittering generalities (virtue)	These are words or symbols in the value system of the target audience that produce a positive image when attached to a person or issue. Peace, hope, happiness, security, wise leadership, and freedom	32
Loaded language	Using specific words and phrases with strong emotional implications (either positive or negative) to influence an audience.	492
Name calling/labeling.	Generate fear and bias using derogatory terms to form an adverse judgment against a person, a group, ideologies, concepts, or institutions that they want us to condemn	288
Obfuscation, intentional vagueness	Using words which are deliberately not clear so that the audience may have its own interpretations	12
Presenting irrelevant data	Introducing irrelevant material to the issue being discussed, so that everyone's attention is diverted away from the points made	1
Repetition	Repeating the same message repeatedly so that the audience will eventually accept it	11
Slogans	A brief and striking phrase that may include labeling and stereotyping. Slogans tend to act as emotional appeals	45
Smears	A smear is an effort to damage or call into question someone's reputation, by propounding negative propaganda	97
Thought terminating	Words or phrases that discourage critical thought and meaningful discussion about a given topic	7
Whataboutism	A technique that attempts to discredit an opponent's position by charging them with hypocrisy without directly disproving their argument	4

Table 2. Our State Backed Propaganda Labeled Data.

Timeframe	Number of states linked accounts	Number of labeled propaganda tweets
December 2019	Saudi_Arabia_112019 (5929 users))	110
October 2020	Saudi Arabia (qatar_082020) (34 users)	108

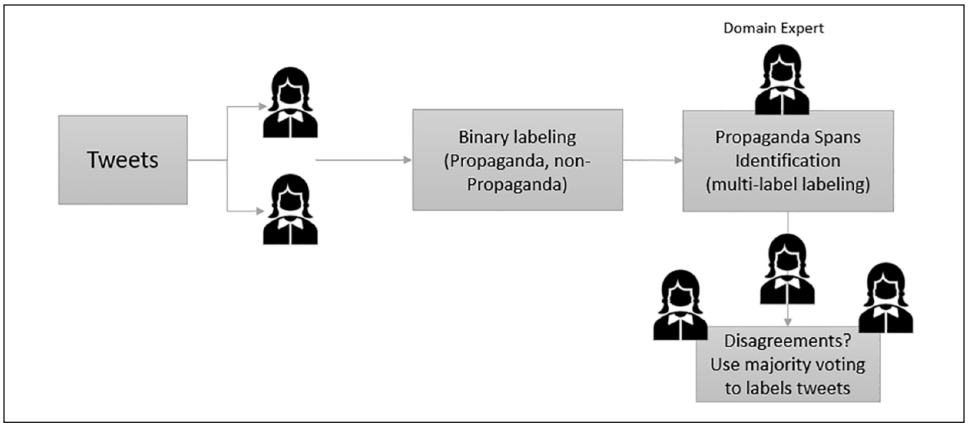

Figure 1. Labeling process of the targeted propaganda data set.

Table 3. Sample of General and Contextualized Propaganda.

Propaganda technique	General propaganda (GP)	State-linked/contextualized and targeted propaganda (TP)
Appeal to authority	An Egyptian observer: The 2014 elections were the declaration of liberation, and the current beginning of reconstruction	America knows that if Saudi Arabia gets angry at it, it means that the entire Islamic world is angry, and only Saudi Arabia may absorb the anger
Appeal to fear/prejudice	The pollution of the largest artificial lakes in Lebanon raises the alarm and warns of an environmental disaster	It is not surprising that Israel and America do not hesitate to punish anyone who opposes them, and there is no consideration for any values or covenants if they contradict their interests, unlike the Arabs
Black-and-white fallacy/ dictatorship	Dialogue—Talal Abu-Ghazaleh: There is no other solution in Palestine except with the end of the occupation	Iran has two solutions, the sweetest of which is bitter; Either being naked for America or being naked for the Iranian people
Causal oversimplification	A bad future awaits humanity. Artificial intelligence is in the dock	The Arab Spring is an American Spring in origin. America has completely laundered its files in the Middle East. It removed agents and brought new agents
Loaded language	UAE Sheikha Jawaher Al Qasimi criticizes the normalization of education with Israel: "Their curricula recommend killing and usurping Arab land.	They will expel his children / Saudi Arabia is the center of Islam and will not allow the dogs of the "Rafidites," their hypocrisy, their ally Israel, and complicit America to celebrate

Table 4. Propaganda Tweets With Propaganda Spans.

Tweet example	Propaganda spans
The trump administration's sanctions on the International Criminal Court deepen America's failure to deliver justice for the most heinous global crimes	"Loaded Language", "Smears"
Syrian regime oil minister: 90% of Syrian oil is under American influence, and its promises of deliveries are fading	"Loaded Language", "Name calling/Labeling"

consists of cases where one of the annotators labeled a tweet as belonging to a propaganda category and the two others disagree. In those cases, the annotators conducted a second round of labeling to avoid labeling unintentional errors; we then used the majority voting rule to produce the final labeled data set. As with the previous research efforts in this area, the distribution of the propaganda techniques in the newly collected data set was skewed. The sample of our labeled contextualized (targeted propaganda) and general propaganda is shown in Table 3. In Table 4, we also show a sample of propaganda spans in the same sentence.

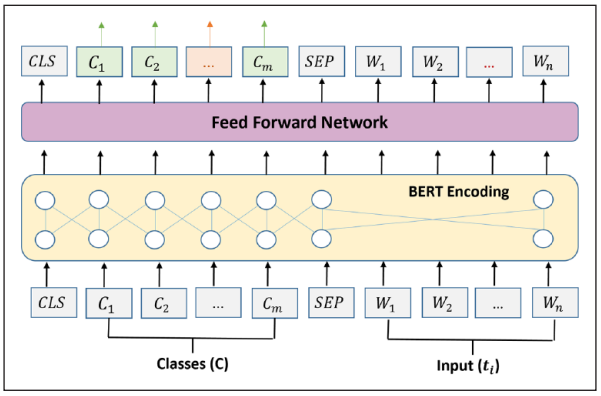

Figure 2. Propaganda detection architecture.

Machine Learning Analysis

We use Alhuzali and Ananiadou's (2021) model, SpanEmo, a machine learning algorithm that can conduct multilabel classification and span detection tasks, to analyze our data[1]. We selected this model for two particular reasons. First, to answer our research question of "To what extent can these issues be described as propagandic in nature?" we need a classification algorithm which can learn from features and labels in the training data and thereafter perform classification tasks on the unseen data (testing data). This task is different from what we commonly see in conventional social science, which is usually causal inference using statistical modeling. Therefore, a machine learning algorithm would be more appropriate to conduct this kind of task. Second, in both of our training and testing data set, each tweet could have more than one label, and this model would allow us to explore the situation where one piece of propaganda uses multiple techniques mentioned before and predict the techniques that are used in the unseen data (testing set). Based on the understanding of multilabeled propagandic tweets, we will be able to distinguish propagandic tweets from nonpropagandic tweets more accurately. Figure 2 summarizes the approach and the deep learning architecture we utilize to detect propaganda categories from the same tweet.

Our propaganda detection approach casts multilabel propaganda classification as span-prediction, which learns associations between labels and words in tweets. Therefore, we utilized the model by Alhuzali and Ananiadou (2021) to classify into 17 categories of propaganda. Let $\{(t_i, y_i)\}_{i=1}^{N}$ be a set of N tweets with the C propaganda classes, where t_i represents the training tweet and $y_i \in \{0,1\}^m$ m are the set of labels. Figure 1 shows how the classes and the training tweets are used as inputs to the model. The training data are processed by a $BERT$ encoding approach developed by Devlin et al. (2018). The inputs to the encoder are the propaganda classes and the training tweets. The hidden layer $(H_i \in R^{T\times D})^2$ containing the training tweets and the set of classes set is obtained as follows: $H_i = Encoder\left(\left[CLS\right]+\left|C\right|+\left[SEP\right]+t_i\right)$. In this formula, $\{[CLS],[SEP]\}$ represents special tokens that are added to the data and $|C|$ denotes the number of propaganda classes. T^2 and D are the length of the input and the dimensionality of data. A feed-forward network (FFN) is utilized, with a nonlinear hidden layer, a Tanh activation $(f_i(H_i))$, and a vector $p_i \in R^D$, which calculates the dot product of f_i and p_i. As our task involved a multilabel propaganda classification, a sigmoid activation is used to determine whether a $class_i$ should be included in the predicted classes as $\hat{y} = sigmoid\left(FFN\left(H_i\right)\right)$. The span-prediction tokens were compared with the ground truth labels as there is a one-to-one mapping with such labels. Using the approach by Yeh et al. (2017). We used the label correlation aware loss as an objective function as $\mathcal{L}_{LCA}\left(y,\hat{y}\right) = \frac{1}{\left|y^0\right|\left|y^1\right|}\sum_{(p,q)\in y^0\times y^1}\exp\left(\hat{y}_p - \hat{y}_q\right)$. This loss function also fits our training objectives to detect co-occurrence of propaganda because it splits labels into positive and negative pairs based on their co-occurrence. In Formula 3, y^0 denotes the set of negative labels and y^1 denotes the set of positive labels. $\hat{y}_p$ represents the p^{th} element of vector $\hat{y}$. The objective of this loss function is to maximize the distance between the labels based on their co-occurrence. In other words, the model loss increases if it predicts a pair of propaganda labels that should not co-exist for a given tweet. As it was the case in Alhuzali and Ananiadou (2021), the model label-correlation loss is combined with the binary cross-entropy. This aims to help the label-correlation loss to focus on maximizing the distance between co-occurrences while at the same time taking advantage of the binary cross-entropy to maximize the probability of the correct prediction. The overall training objective was computed as follows $\mathcal{L} = \left(1-\alpha\right)\mathcal{L}_{BCE} + \alpha\sum_{i=1}^{M}\mathcal{L}_{LCA}$, where $\alpha \in [0,1]$ denotes the weight used to control of each loss function.

Experiments

Using both data sets, we conducted three types of experiments on our deep learning model:

Table 5. Hyperparameter Settings.

Parameter	Value
Feature dimensions	732
Batch size	6
Early stop patience	50
Number of epochs	50
lr-BERT	2e-5
Optimizer	Adam
Alpha α	0.2
Drop out	0.1

- GP: The training and development data were selected from the General Propaganda (GP) Data Set 1. The testing data are selected from our Targeted Propaganda (TP) Data Set 2.
- TP: The training, development, and testing data were selected from our TP Data Set 2.
- FT: Training data are selected from GP; then, the model is finetuned (FT) on the TP data from Data Set 2.

All our experiments were conducted using the same hyper parameters settings shown in Table 5.

The technical settings of our experiments consist of using the PyTorch implementation (Paszke et al., 2017), HuggingFace implementation of BERTBASE (Wolf et al., 2020), and BERTBASE-ARABIC (Safaya et al., 2020) as a pretrained Arabic-language model. We leave testing our approach on other Arabic-language models such as ARABERT and MARBERT as a future work. Table 6 shows the characteristics of Training (T), Development (T), and Testing (TS) experimental data selected from both GP and TP data sets. Several evaluation measures were used in our experiments as follows:

- The Jaccard Similarity (JaccS): This measure is usually used to quantify the degree of similarity between two sets, such as a set of true labels and a set of predicted labels. It is calculated by dividing the size of the intersection of these two sets by the size of their union. The Jaccard Similarity measures how much overlap or commonality exists between the elements present in the true label set and the predicted label set. A higher Jaccard Similarity score indicates a greater degree of overlap or agreement between the two sets, while a lower score suggests less agreement and more dissimilarity. As we are measuring the reliability of classification for a multilabel classification, which involves assigning multiple labels or categories to each instance, the Jaccard Similarity handles this scenario making it well suited for evaluating the overlap between predicted labels and true labels.
- Macro and Micro F1: Macro-F1 is calculated by averaging the F1 scores for each class individually, while micro-F1 is calculated by averaging the precision and

Table 6. Training, Development, and Testing Data Sets.

Technique	GP			TP			FT		
	T	D	TS	T	D	TS	T	D	TS
Appeal to authority	21	7	2	2	1	2	21	1	2
Appeal to fear/prejudice	48	7	1	3	1	1	48	1	1
Black-and-white fallacy/dictatorship	2	1	1	2	1	1	2	1	1
Causal oversimplification	4	1	1	1	1	1	4	1	1
Doubt	29	1	7	29	18	7	29	18	7
Exaggeration/minimization	44	10	1	8	2	1	44	2	1
Flag-waving	5	2	5	11	7	5	5	7	5
Glittering generalities (virtue)	25	7	2	2	3	2	25	3	2
Loaded language	446	46	6	6	5	6	446	5	6
Name calling/labeling.	244	44	8	3	4	8	244	4	8
Obfuscation, intentional vagueness	9	3	3	7	7	3	9	7	3
Presenting irrelevant data (red herring)	1	0	4	4	3	4	1	3	4
Repetition	9	2	1	1	1	1	9	1	1
Slogans	44	1	1	1	1	1	44	1	1
Smears	85	12	2	6	5	2	85	5	2
Thought-terminating cliché	6	1	2	2	1	2	6	1	2
Whataboutism	3	1	1	4	1	1	3	1	1

Note. GP = general propaganda; TP = targeted propaganda; FT = finetuned; TS = testing.

recall scores for all classes, regardless of their size. Macro-F1 is used when some classes are more important than others, or when we have a data set that is balanced. Micro-F1 is a good metric to use when all classes are equally important, or when we have a data set that is imbalanced. In this work, we used both measures to get a more complete picture of the performance of their model on both balanced and imbalanced data sets.

- Training loss: It is a metric that quantifies how well a machine learning model is performing on its training data during the training process. It represents the error or the difference between the predicted values and the actual target values for the training examples.
- Validation loss: It is a metric used to evaluate a machine learning model's performance on data that it has not seen during training. This separate data set, called the validation set, is distinct from the training data.

Results

The results of our experiments are shown in Figures 3 and 4. Figure 3 shows our experimental results using the Saudi accounts tweets as our TP data. We noticed significantly better metrics when training using TP data compared with training using GP (Figure 3b compared with a). We run our experiments on micro F1-score, macro F1-score, and Jaccard index score. The latter is the size of the intersection divided by the size of the union of the true label set and predicted label set. We observed better F1-Micro, F1-Macro, and JS using training on TP. We also observed that training on GP and then finetuning the model using TP data are the least effective approach to detect targeted propaganda (Figure 3c). This result was validated on the testing data as well, where training using TP led to the best model accuracy in terms of our classification metrics (Figure 3d). It is also noticed that training and validation using TP leads to the most stable model in terms of both training and validation losses (Figure 3b). Results on Qatari accounts tied to the Saudi government were not significantly different compared with Saudi accounts (Figure 4). It is noticed, however, that FT leads to relatively better results compared with Saudi accounts (Figure 4c compared with 3c). We think that this is related to the sources where the GP data set was collected, where it was mainly from some Qatari news networks such as Qatar's Aljazeera news network. As such, similarity between TP of Qatari accounts and GP data leads to better finetuning results.

The results shown in Figure 3a and b highlight the advantages of incorporating targeted propaganda for training purposes. Specifically, as seen in Figure 3b, the F-Micro score achieves a noteworthy increase, reaching 0.63, compared with 0.56 in scenarios employing general propaganda data (Figure 3a). In addition, the Macro F1 scores exhibit significant improvement when the model is trained using the targeted propaganda data set (as demonstrated in Figure 3b). Moreover, training and validation losses highlight the model's robust convergence when trained with the targeted propaganda data set. Figure 3b illustrates a gradual reduction in both training and validation loss values, compared with Figure 3a, where the model's learning capabilities are less effective. It is important to note that an alternative approach,

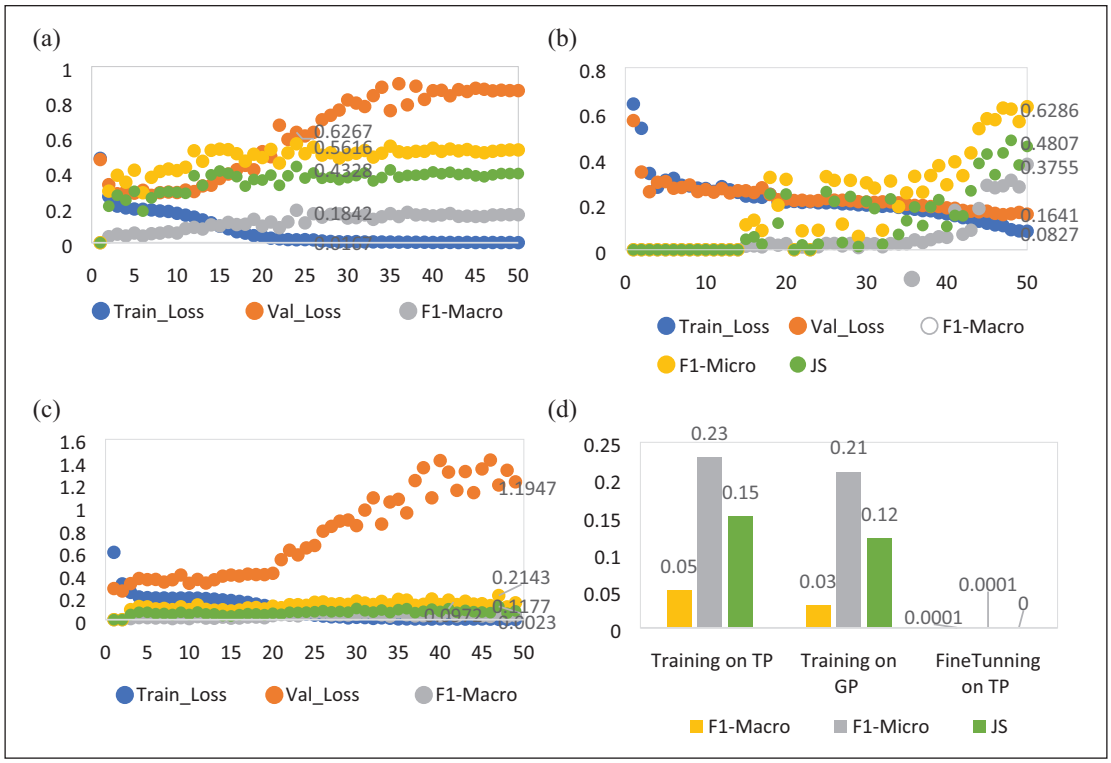

Figure 3. (a) Training using GP. (b) Training using TP. (c) FT on TP. (d) Results on testing data.

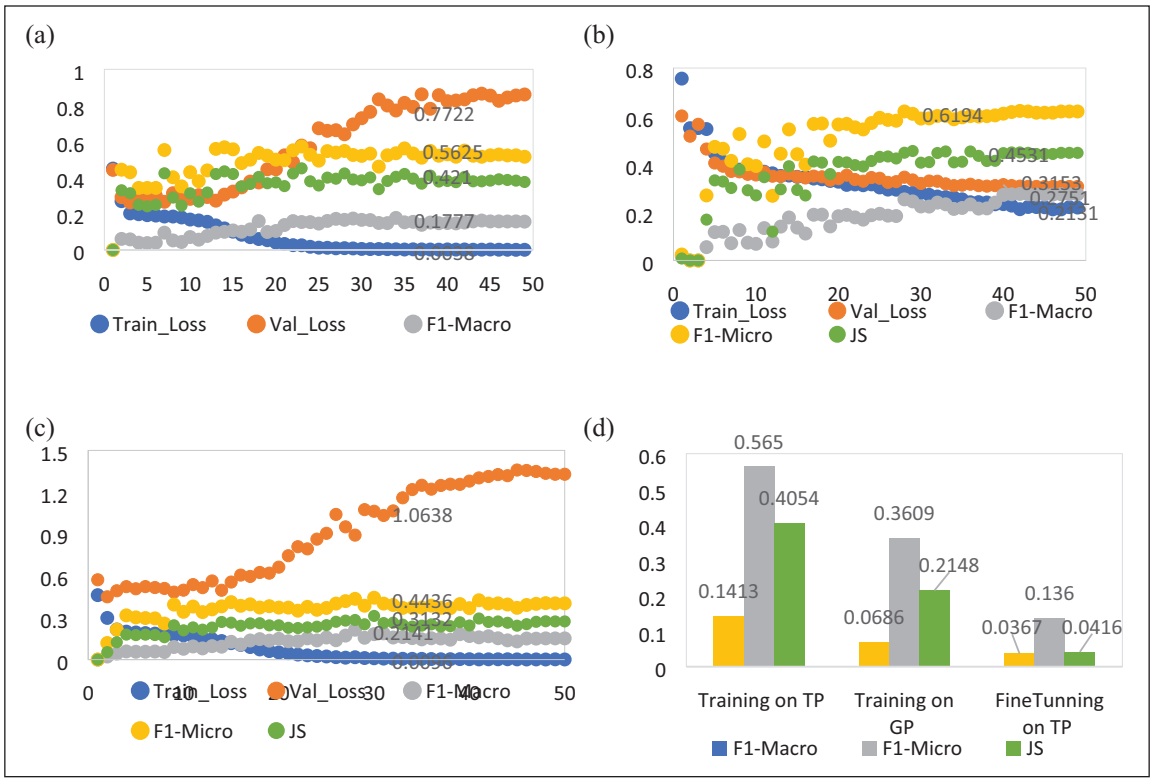

Figure 4. (a) Training using GP. (b) Training using TP. (c) FT on TP. (d) Results on testing data.

involving initial training with general propaganda data followed by finetuning with targeted propaganda data, results in less significant results. Specifically, this approach yields a low JS (Jaccard Similarity) score of approximately 0.11, along with F1-Micro and F1-Macro scores of 0.21 and 0.10, respectively. Furthermore, the consistency of these results on the testing data set, as depicted in Figure 3d, shows the advantages of incorporating targeted propaganda data when predicting targeted propaganda attacks. Figures 4a to d shows almost similar results patterns.

The Case of Saudi Arabia

To illustrate this article's points more specifically and contextually, a single case study is needed to illustrate better the issues discovered. To delve more deeply into the topics, this article presents a case analysis of Saudi Arabia. Before we get to case-specific results on Saudi Arabia, it is important to place the case in the context of this case study. There are several reasons to investigate propaganda activities in Saudi Arabia as opposed to the other countries also investigated in the article, which includes its significant role in geopolitical events. First, it should be noted that Saudi Arabia has one of the highest percentages of social media users in the Middle East (25 million people out of a total population of just under 35 million, and within this, there are an estimated 20 million Twitter users in Saudi Arabia; Shakil et al., 2021). In addition, as a major energy producer Saudi Arabia plays a vital role in the global energy market. In fact, Saudi Arabia provides around 15% of the U.S. crude oil imports and more than 15% of the global crude oil imports yearly between 2000 and 2017 (Beckman & Nigatu, 2021). Furthermore, research shows that negative oil-supply shocks due to sanctions, wars, policies, or natural disasters in Saudi Arabia have an immediate and permanent increase in global oil prices as other oil producers cannot make up for the decrease in Saudi Arabian oil production (Mohaddes & Pesaran, 2016). Saudi Arabia holds a unique religious significance in the Muslim world. Saudi Arabia is the home of Mecca and Medina, the two holiest cities in the Muslim World, and the state "continues to be a center of religious training and its soft power influence in the Sunni World is unmatched" (Ciftci & Tezcür, 2016, p. 6). In addition, although Saudi Arabia may not be willing to openly endanger its relationship with the United States, recent evidence shows that Saudi Arabia has been involved in influence operations against the United States on social media. For example, two Twitter employees were charged with spying for Saudi Arabia on U.S. soil in November 2019 (Barrie & Siegel, 2021), and hashtags such as "Agents of the Embassies" have been used by the Saudi state on social media to resist influence from the Western world (Abrahams & Leber, 2021a). Therefore, we decided to focus on Saudi Arabia as one of, if not the most important great power within the region as it pertains to U.S. interests and great-power competition.

In fact, David Long (2019) argues that there is a unique relationship between Saudi Arabia among the Arab world toward the United States as it holds constant the need for close relations with the United States. In addition, with current tensions rising between Saudi Arabia and Iran, Israel's status in the Middle East (Beck, 2020), and the continuing crisis in Yemen (Darwich, 2020) where Saudi intervention has played a vital role, Saudi Arabia makes a great choice as a case study to examine in this context. It is important to note that as social conditions in Saudi Arabia have been progressing, there have been challenges to the regime; to prevent widespread opposition, the government has significant control over the internet and its censorship powers are steadily increasing (Chaudhry, 2014). Even with censorship, however, there have been movements in the Kingdom's Twittersphere promoting change through hashtag campaigns, for instance, including #women2drive (Chaudhry, 2014).

Abrahams and Leber (2021b) note that Saudi Arabia is one of the most prolific authoritarian regimes in the Middle East that use Twitter as a form of control and as a source of power, to the point that the regime managed to even place spies within the company itself. They note that much of the pro-authoritarian speech on Saudi Twitter is the result of organic activity driven by influential accounts that have built up their followings by toeing the party line—either voluntarily or by regime pressure (Abrahams & Leber, 2021b, p. 1174). Other research concerning Saudi Arabia has focused on emotional analysis of Twitter users living in holy cities compared with more secular metropolitan centers (Shakil et al., 2021), and engagement of Saudi citizens with IO campaigns in general compared with more mainstream news outlets (Barrie & Siegel, 2021). In fact, these authors find that engagement with IO within the Kingdom is not substantial compared with the level of Twitter uses overall, even during significant news events such as the murder of Jamal Khashoggi (Barrie & Siegel, 2021). Now that some context has been provided, we turn to our specific findings.

We analyze propagandic tweets in Data Set 2 from Saudi Arabia using the BEND social cyber security framework proposed by Carley (2020). The BEND framework argues that "influence campaigns are comprised of sets of narrative and structural maneuvers, carried out by one or more actors by engaging others in the cyber environment with the intent of altering topic-oriented communities and the position of actors within these communities" (Carley, 2020, pp. 371–372). In other words, the BEND framework considers both the narratives in the cyber environment and the social networks of cyber actors. This framework explores communication objectives from two dimensions: whether the objective is positive or negative, and whether the objective is aimed at manipulating the narrative (positively or negatively) or manipulating the social networks (positively or negatively). This gives us a two-by-two table in which 16 maneuvers are identified, and the details are shown in Table 6. The BEND framework

Table 7. The BEND Framework.

	Manipulating the narrative		Manipulating the social network	
Positive	Engage	Messages that bring up a related but relevant topic	Back	Actions that increase the importance of the opinion leader or create a new opinion leader
	Explain	Messages that provide details on or elaborate the topic	Build	Actions that create a group or the appearance of a group
	Excite	Messages that elicit a positive emotion such as joy or excitement	Bridge	Actions that build a connection between two or more groups
	Enhance	Messages that encourage the topic-group to continue with the topic	Boost	Actions that grow the size of the group or make it appear that it has grown
Negative	Dismiss	Messages about why the topic is not important	Neutralize	Actions decrease the importance of the opinion leader
	Distort	Messages that alter the main message of the topic	Nuke	Actions that lead to a group being dismantled or breaking up, or appearing to be broken up
	Dismay	Messages that elicit a negative emotion such as sadness or anger	Narrow	Actions that lead to a group becoming sequestered from other groups or marginalized
	Distract	Discussion about a totally different topic and irrelevant	Neglect	Actions that reduce the size of the group or make it appear that the group has grown smaller

Source. Carley (2020).

(Table 7) has been adopted to explore social media content and networks relating to COVID-19 vaccinations (Blane et al., 2022, 2023). Furthermore, Danaditya et al. (2022) have also applied the BEND framework to analyze Indonesian Twitter content, another Muslim community, and by analyzing the maneuvers of narrative and network, they find that a small group of coordinated agents can lead to polarization in the online sphere. Therefore, we believe that the BEND (Carley 2020) framework can provide more insights on the analysis of the Saudi Arabia Twitter sphere.

We conducted an annotation process to map each of the propaganda categories to one or more of the BEND communication objectives. During this phase, we selected three annotators, two of them were media experts who speak Arabic and English. The third expert was an IT professor who has expertise in Natural Language Processing. The annotators, first, reviewed the descriptions of various propaganda techniques. They also reviewed the descriptions of communication effects as defined by the BEND framework. The annotators then proceeded to map each propaganda technique to one or more communication effects. This step involves making associations between techniques and their expected effects. To validate our mapping, we asked annotators to find at least one example from the data set that supports their annotations. This practical validation step is crucial for assessing the real-world applicability and accuracy of the mapping. As the size of the targeted propaganda data is small, the annotators focused partially on a subset of 11 communication objectives that were selected from both manipulation strategies and that are clearly presented in the targeted propaganda data set. While we didn't directly measure the impact on the social network, we rely on the narratives that may lead to such an impact. Table 8 shows one of those examples. For this labeling activity, we asked the annotators to resolve any disagreement between them, through comprehensive discussion among them.

Table 8. Mapping Between Targeted Propaganda and BEND Communication Effects.

Example:
The Arab Spring is an American Spring in origin. America has completely laundered its files in the Middle East. It removed agents and brought new agents.
Strongest Propaganda Span:
Causal Oversimplification
Communication Effect: "Dismiss"
The statement simplifies complex geopolitical events like the Arab Spring by suggesting a direct cause-and-effect relationship between the Arab Spring and American involvement. It implies that the Arab Spring is solely an outcome of American actions, which is an oversimplified view of a multifaceted situation. This oversimplification can lead to the "Dismiss" effect by downplaying other complexities of the Arab Spring.

The mapping results are shown in Figure 5. Some BEND strategies are associated with more than one propaganda technique, for instance, the excite effect is associated with six types of propaganda.

A frequency analysis of mapping the propaganda found in the state-linked data about the Saudi Arabia data set is shown in Figure 6. As most of the analyzed propaganda is aimed at criticizing, undermining, or discrediting the United States, "Dismay" and "Neutralize" strategies were frequently used. "Dismay" strategies aim to elicit fear, worry, or doubt, which can be effective in making an audience question or reject the target of the propaganda. "Neutralize" strategies, on the contrary, aim to diminish or dismiss the impact of an opposing viewpoint, which can be useful in a setting where there are conflicting ideas or narratives. As the effectiveness of a

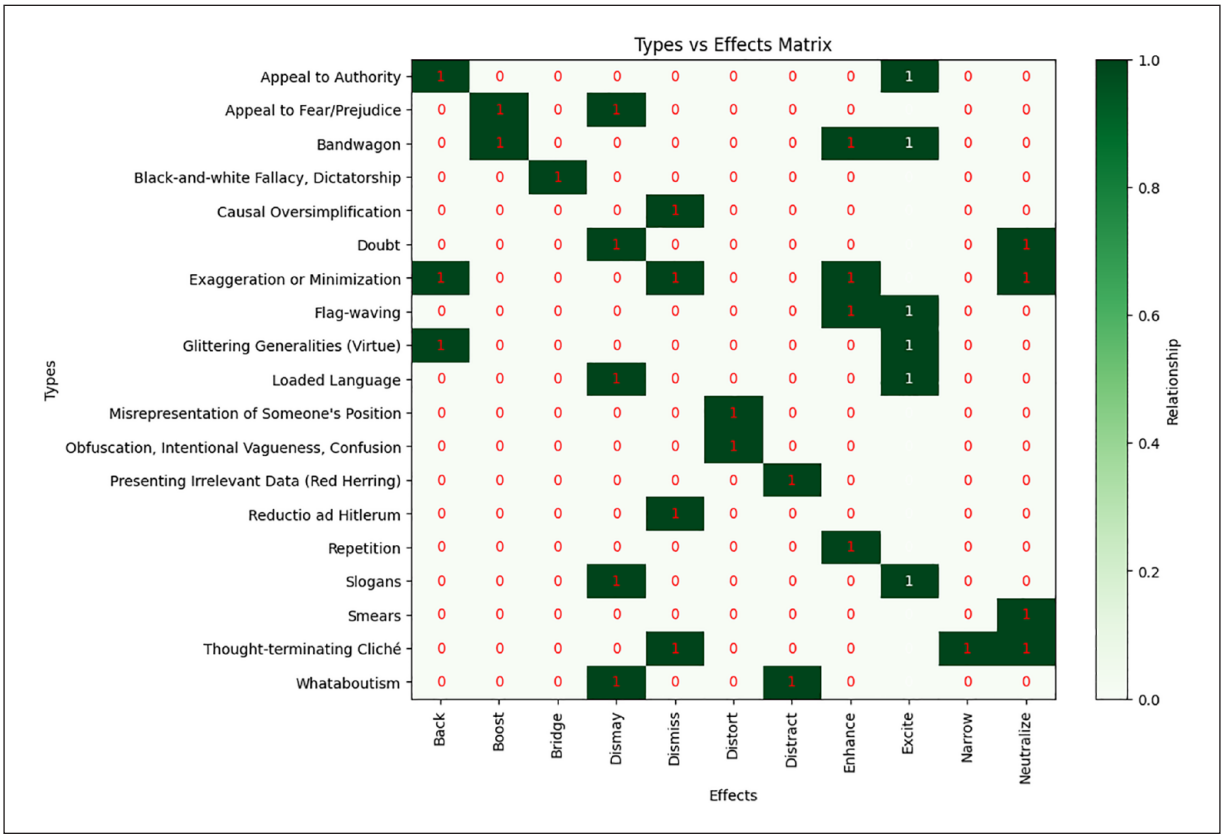

Figure 5. Mapping between propaganda categories and the BEND framework communication objectives.

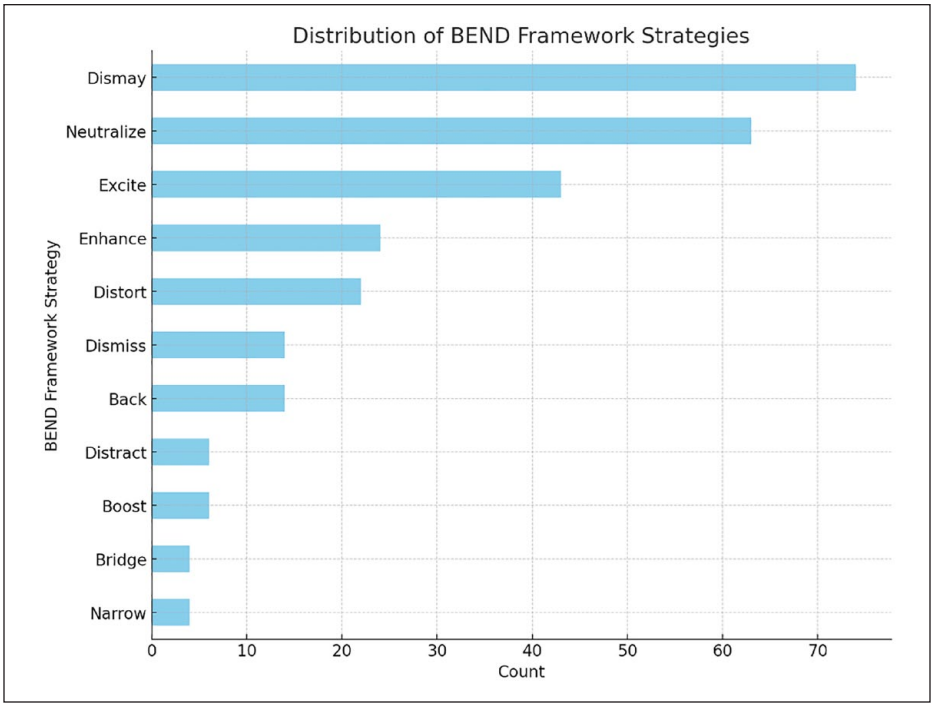

Figure 6. Frequency analysis of BEND communication strategies.

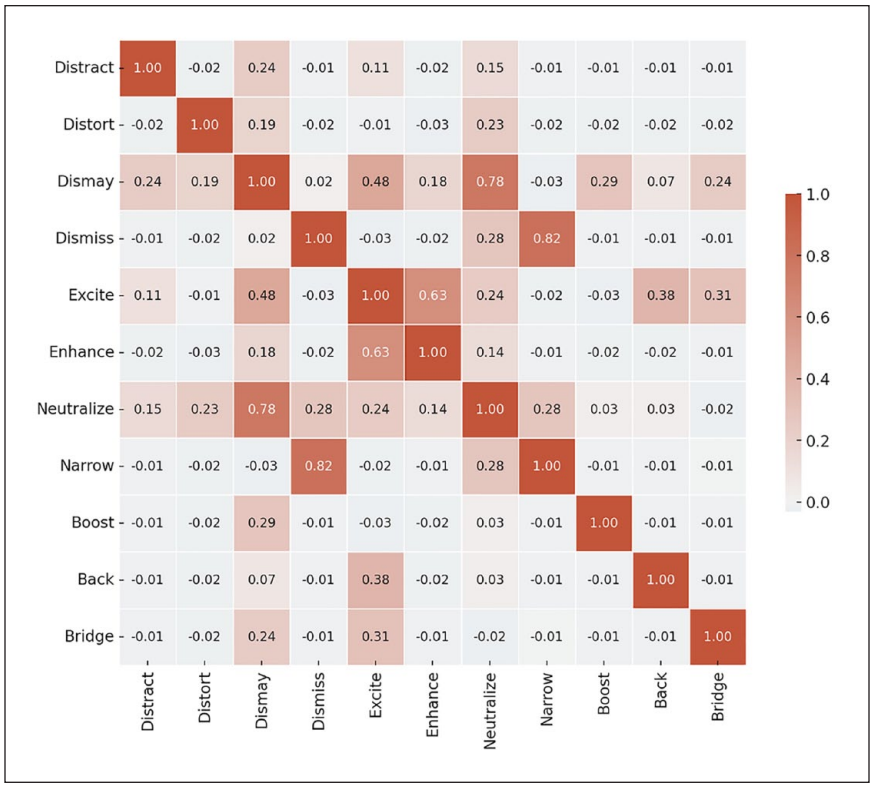

Figure 7. Correlation between BEND communication strategies in the Saudi propaganda data.

propaganda technique often depends on how it resonates with its intended audience, "Dismay" and "Neutralize" strategies are particularly effective at influencing the attitudes and behaviors of the audience with respect to U.S. politics. The overall goals of the propaganda campaigns in state-linked data set also influence the choice of strategies. As the goal was to create confusion or stoke fears about U.S. policies, the "Dismay" strategy was a logical choice. The goal to challenge or undermine an opposing viewpoint also justifies the neutralize strategy.

We then conducted a correlation analysis to examine which strategies are used together. The correlation analysis results shown in Figure 7 also support our analysis. We can draw several findings from this correlation study as follows:

- Dismiss and Enhance (correlation: –0.02): The techniques of dismissal and enhancement do not often appear together in the analyzed propaganda data set. This suggests that Saudi state-linked actors dismiss certain information (likely negative information about Saudi Arabia or its policies); they do not typically try to enhance or amplify other information (e.g., positive information about Saudi–U.S. relations) within the same message.
- Dismay and Neutralize (correlation: 0.78): When trying to create a sense of fear or concern among audiences (Dismay), these actors often aim to neutralize some viewpoints. This means that they are trying to induce worry about certain topics while simultaneously reducing the impact or credibility of opposing viewpoints by the United States. Some supporting statements from the data set is "America will soon face its own trials and will perish. We seek refuge from its evil. InshaAllah." This statement expresses dismay toward America, suggesting that it will face negative consequences. The mention of seeking refuge from its evil attempts to neutralize the power or influence of America. Another statement is "The Iraqi government is elected!? I remember Ayad Allawi won the elections, but Iran and America want him destroyed." This example expresses dismay toward the Iraqi government, suggesting that it is not truly elected and controlled by external forces (Iran and America). By highlighting the influence of Iran and America, the statement attempts to neutralize the legitimacy or credibility of the Iraqi government.
- Dismay and Excite (correlation: 0.48): The positive correlation suggests that messages often aim to both arouse alarm and strong emotions or enthusiasm among audiences. This can be seen as strong emotions that favor the Saudi narrative while creating concern or fear about alternative viewpoints or actions by the United States.

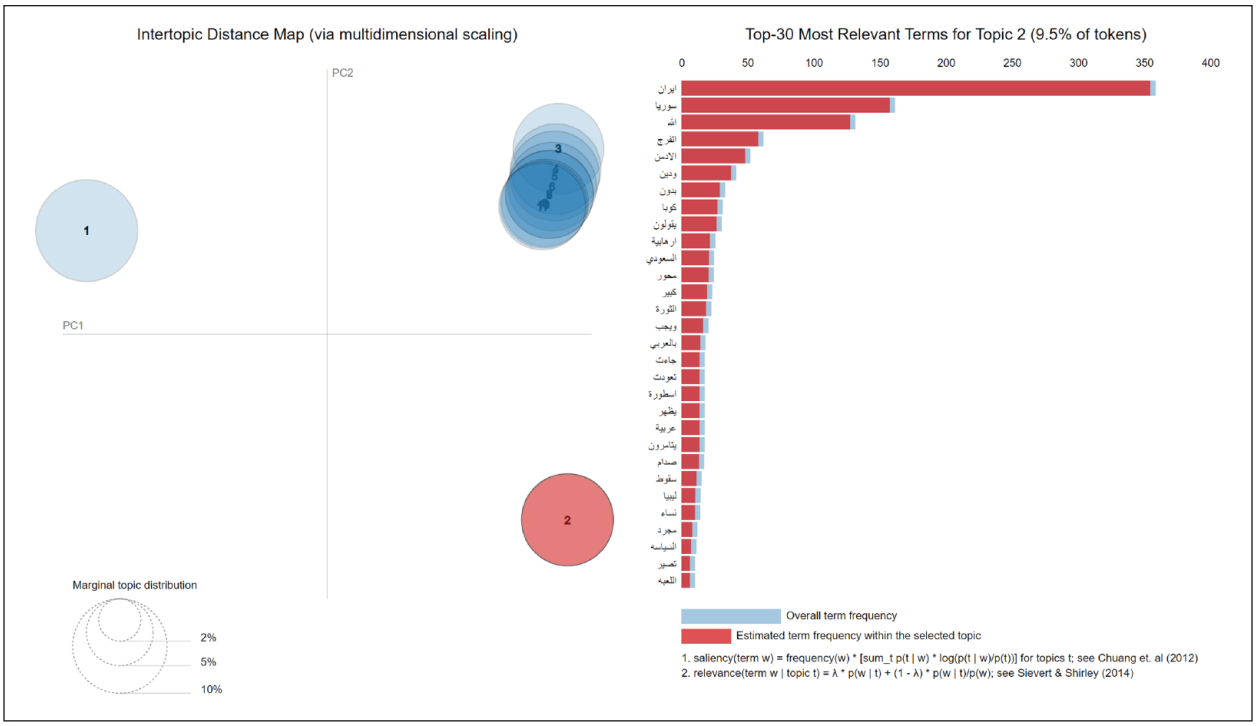

Figure 8. Topic analysis: Thematic Topics Sample 1.

- Excite and Enhance (correlation: 0.63): When Saudi state-linked actors aim to stir up strong emotions, they often try to amplify or enhance certain viewpoints or pieces of information. This could be part of a strategy to create emotional attachment to the messages they want to promote, such as positive perceptions of what Saudis did to the United States.
- Neutralize and Narrow (correlation: 0.28): This correlation indicates that attempts to neutralize certain viewpoints (likely those opposing Saudi interests) often go hand in hand with narrowing the range of debate. This could be a strategy to control the narrative within the United States, neutralizing opposing viewpoints, and limiting the discussion to topics where Saudi state-linked actors can more effectively push their preferred narrative.
- Dismay and Neutralize: Because this article focuses on targeted propaganda operations, our explanation of strong correlation between Dismay and Neutralize is that both techniques aim at creating a sense of doubt, or confusion among the audience. This contrasts with "Excite," which aims to invigorate or energize the audience. Both "Dismay" and "Neutralize" aim to undermine the opponent's message, thereby making them more compatible. Finally, "Dismay" may create an emotional state in the audience that makes them more susceptible to "Neutralize" tactics, which could aim to diminish the credibility or importance of opposing views.

In the course of our research, we implemented a thematic analysis on the provided data sets, utilizing Latent Dirichlet Allocation (LDA) to identify prevalent topics. Our topic analysis results are shown in Figures 8 and 9. As shown in Figure 8, a subset of the anti-U.S. propaganda was notably linked to terms such as "IRAN," "SYRIA," "SHIA," "ALEPPO," and "SADDAM" (translation from English to Arabic is provided in Table 9). Upon further examination of tweets containing these terms, it became evident that the associated user accounts were mainly critiquing U.S. policies in Syria, Iraq, and Yemen. A belief held by some Saudis is that the United States is unfairly favoring Iran-backed Houthi rebels in Yemen. These individuals contend that while the United States has expressed criticism of Saudi Arabia's military intervention, it has not sufficiently held the Houthis responsible for their deeds. Such perceived bias incites unfavorable attitudes toward U.S. politics. For instance, several of the tweets analyzed for this article demonstrate this precise belief. One user commented, "Americans love parasites that spread poison into countries and destroy their stability if you notice that America did not harm the Houthis, Hezbollah, and Muqtada al-Sadr." Another user commented, "The Houthis display slogans against America while it supports them and they support it! Are they America's hand at Saudi Arabia's side so that Saudi Arabia can blackmail them whenever it wants?!" State-linked accounts also have negative views about the U.S. role in Syria. Given the complex regional dynamics, some Saudis may perceive U.S. policies in Syria such as engaging the Iranian government in

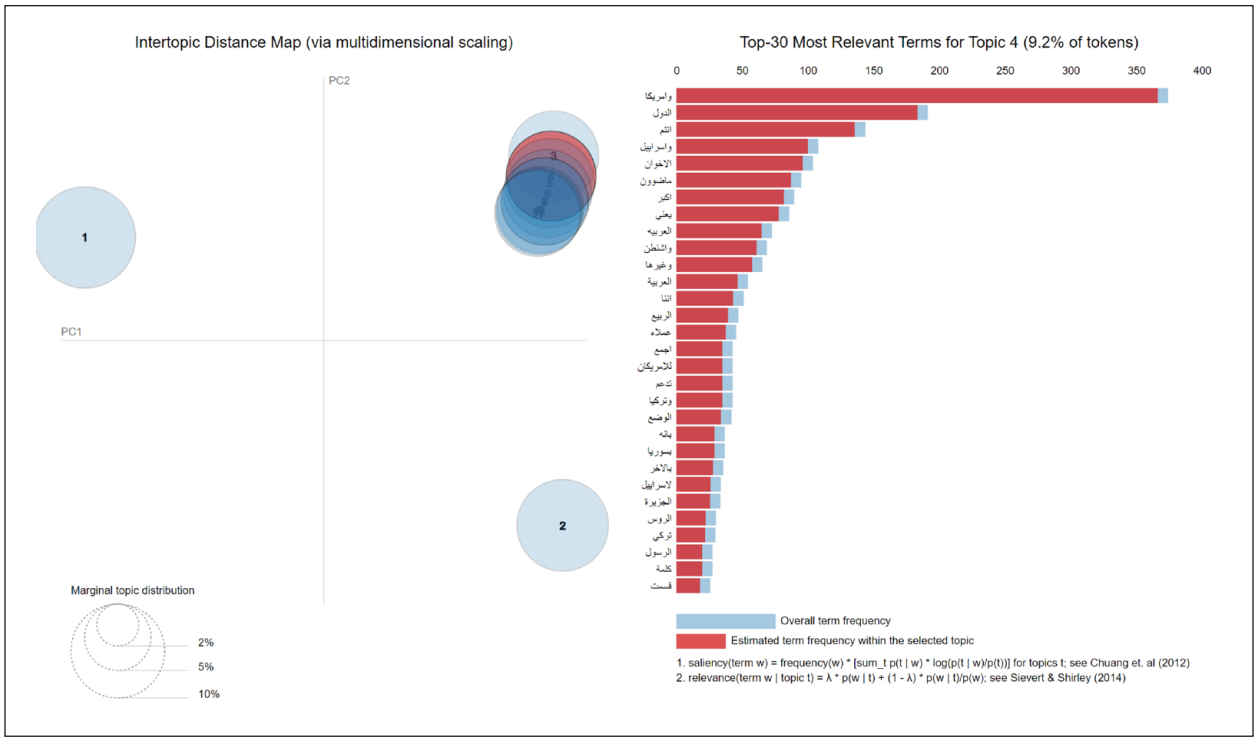

Figure 9. Topic analysis: Thematic Topic Sample 2.

Table 9. Translation of the Most Important Keywords in Figure 8.

Iran	Syria	Allah	terrorist	Saudi	ally	revolution	legend	Arabic	They conspire	Saddam	Fall	Women	Politics	Game
إيران	سوريا	الله	ارهابية	السعودي	محور	ثورة	أسطورة	عربية	يتكمرون	صدام	سقوط	النساء	السياسة	لعبة

Table 10. Translation of the Most Important Keywords in Figure 9.

USA	Israel	Muslim Brotherhood	Washington	Arab Spring	Support	Turkey	Syria	Russian	Prophet	Going on	Divided
امريكا	اسرائيل	الإخوان	واشنطن	الربيع	تدعم	تركيا	سوريا	الروس	الرسول	ماضون	قسمت

regional diplomacy over the Syrian civil war as indirectly benefiting Iran (Gause, 2016). They believe that U.S. actions have allowed Iran to expand.

As shown in Figure 9 and the translation in Table 10, we identified topics pertaining to matters such as the Muslim Brotherhood. Perceived inconsistencies in U.S. policies have led to the belief that the United States endorses the Muslim Brotherhood. For instance, in the context of the Arab Spring uprisings, the United States seemed to initially back certain Islamist groups linked to the Muslim Brotherhood. This has contributed to the impression that the United States holds a favorable stance toward this organization.

Relatedly, most of the discourse about Russia between 2019-2021 focuses on the Russian role in Syria. The role that Russia has played in the Syrian conflict, supporting the Syrian government, has made some Saudis recognize Russia's growing influence in the region and adjusted their discourse, accordingly, acknowledging the need for engagement and dialogue with Russia (Suchkov, 2016). We collected a sample of tweets in 2022 and found that most of the Saudi influencers on Twitter are supporting reforming the Saudi–Russian relationship. Figure 10 shows a sample of a topic analysis of such tweets (translation from English to Arabic is provided in Table 11). Most of the tweets show that influencers focused on "creating a balanced relationship with Russia which should have been invested in decades ago where Saudis sacrificed for the sake of Americans." Many actors believe that if the relationship with Russia had been balanced, Saudi Arabia would have a nuclear program.

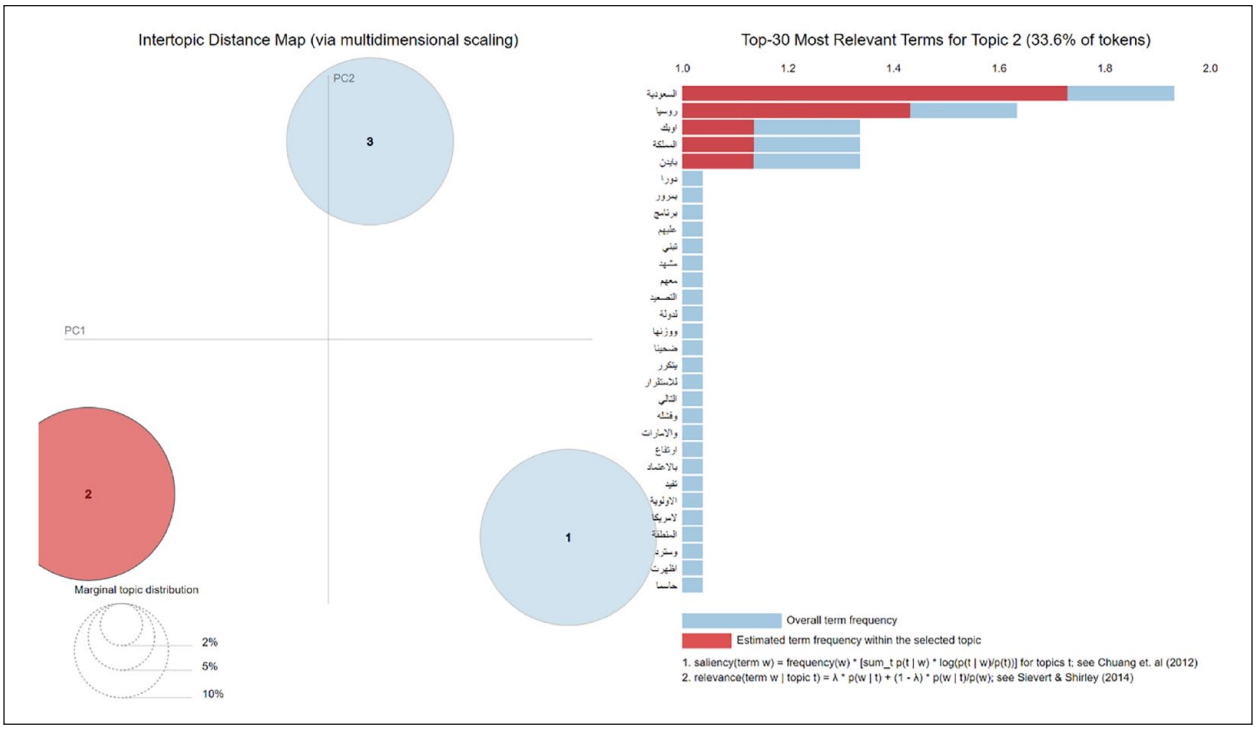

Figure 10. Topic analysis: discourse shifts toward Russia.

Table 11. Translation of the Most Important Keywords in Figure 10.

Saudi Arabia	Russia	OPEC	Biden	Program	USA	UAE	Turkey	Scene	Weight	High	definitive	sacrifice
السعودية	روسيا	اوبك	بايدن	برنامج	امريكا	الامارات	تركيا	مشهد	وزن	ارتفاع	حاسما	تضحية

Note. OPEC = Organization of the Petroleum Exporting Countries.

The analysis above shows that most of the state-linked activities in the analyzed data set are related to state-linked accounts connected to the government in Saudi Arabia. However, that does not exclude efforts by a Russian agent on social media. Those activities often include elements, including propaganda, disinformation, and targeted messaging. Russian media outlets highlight instances where the United States has taken actions contrary to Saudi interests or where policy decisions appear inconsistent. By emphasizing these instances, Russian propaganda attempts to create doubt and erode trust in the U.S.–Saudi alliance, presenting Russia as a more consistent and dependable partner.

U.S. Policy Recommendations

Any attempt from democratic regimes to engage in propaganda or counter propaganda emanating from authoritarian regimes like Saudi Arabia could be seen as dangerous to democracy itself, and it rife with potential problems. To not endanger democracy—a recurring theme right now within the United States' domestic sphere's nexus with national security (Schünemann, 2022), several policies are recommended on how to combat propaganda and influence operations that fall within democratic norms the United States and other democracies can follow. First, as Woolley (2022) notes, "Governments and other institutions working to push back against the cascade of digital falsehood . . . must be clear about the values that drive manipulation campaigns, particularly when autocrats are behind them" (p. 127). In other words, for the United States to effectively combat Saudi propaganda and influence operations more generally, it must first be transparent about what it is doing and why it is doing so.

When engaging in counter-propaganda activities, there must be given clear reasons as to why they are needed; furthermore, the reasons why, in this case, Saudi Arabia targets the United States must be explained within official settings, particularly Congressional hearings, thereby easing the democratic impulse to not engage in such operations as well. It must be made clear to the public, through these types of hearings, what are the mission-set, target vectors, strategy, motivations, and tactics behind Saudi campaigns and how the

United States will respond to them. Of course, for national security purposes, nothing classified needs to be divulged in this context. However, there should be some explanation of these campaigns in perhaps the National Cyber Strategy, or National Defense Policy. Next, and building further on Woolley, social media firms within the United States must do a better job at managing propagandic and influence operations emanating from external nation-state actors; in other words, even if the content is allowed to stay, or accounts are not blocked, it is necessary to publish these events to the public in an attempt at democratic transparency (Woolley, 2022, p. 127). Both Meta and Twitter already do this in some respects, but the data and reports do not seem to be readily available to the general public. Social media firms making these self-studies more accessible to is a step in the right direction toward democratic citizens understanding the online threat derived from propagandic influence operations.

Third, as Albert et al. (2023) argue, for the United States to counter Information Warfare and Influence Operations (IWIO), of which propaganda is a tool, perhaps the most important step is for the United States to have a whole-of-government approach. This could entail creating a new entity that handles all influence operations online, or, if it entails creating something similar to a Joint Operations Command, which would unify all entities within the government concerning doctrine, strategy, and intelligence related to countering propaganda (Albert et al., 2023). What is clear is that the United States lacks a unified doctrine for influence operations, and this puts the United States at a disadvantage, especially against more unified, autocratic regimes, especially those such as Russia and China, but theoretically, Saudi Arabia as well. As it stands, the United States is lagging behind more authoritarian adversaries in the information domain generally because of its "inability to turn data into operational intelligence and its lack of human capital allocation regarding [information warfare] IW" (Albert et al., 2023). The present research demonstrates the parameters in which propaganda operations targeting the United States exist, and thus, help set the operating principles for a whole-of-government approach the United States would need to effectively and ethically counter propaganda, within democratic norms.

Conclusion

Social media threats are becoming increasingly like conventional cyberattacks. In areas such as social engineering, attacks are targeting specific individuals or organizations. This is also true in information operations such as propaganda campaigns that target certain countries, individuals, and organizations. We collected targeted propaganda spans of anti-U.S. texts from different geopolitical contexts and showed that the general frameworks to detect those types of attacks are not effective. We show that social cyber-attack detection models need to be contextualized, meaning that if they target specific groups, countries, or organizations, model finetuning approaches may not be sufficient to identify what propaganda effects attackers are aiming to achieve. This has two implications: Theoretically, it implicates reconsidering the existing behavioral models of the social cyber-attack intentions in low resource languages such as Arabic: for example, considering an extension of the BEND Social-Cyber security framework (Carley, 2020). Practically, our research implicates the need to consider semantics and context to detect those attacks effectively.

We introduced one of the first research attempts to investigate contextualized state-backed social media attacks in the Middle East. We used general training models to detect political propaganda on U.S. personnel and institutions. Our results indicated the limitation of the general propaganda detection models to identify more targeted forms of propaganda. We recommend a possible extension of the existing classification of social cyber threats in other languages such as Arabic. We believe that there is a need for a new sociotechnical framework to detect such attacks, and the authors of this article are currently researching this. Specifically, we will create a revised BEND framework for Arabic social media. As an extension of this work, we believe that we also need to increase the size of the targeted data set as one of the limitations of this work. We also consider studying the emotional reactions of the targeted propaganda attacks. Finally, we will test our method on other Arabic-language models.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This material is based upon research supported by the U.S. Office of Naval Research under award number N000142212549.

ORCID iDs

Craig Douglas Albert https://orcid.org/0000-0003-3225-9386
Ahmed Aleroud https://orcid.org/0000-0003-4337-1488

Note

1. The implementation is an extension of the span detection implementation provided by Alhuzali and Ananiadou (2021), which can be found in the following repository: https://github.com/hasanhuz/SpanEmo. The data sets and code are available upon request to replicate experiments.

References

Abrahams, A., & Leber, A. (2021a). Comparative approaches to mis/disinformation| electronic armies or cyber knights? The sources of pro-authoritarian discourse on Middle East Twitter. *International Journal of Communication*, *15*, Article 27.

Abrahams, A., & Leber, A. (2021b). Electronic armies or cyber knights? The sources of pro-authoritarian discourse on Middle East Twitter. *International Journal of Communication*, *15*, 1173–1199.

Alam, F., Mubarak, H., Zaghouani, W., Martino, G. D. S., & Nakov, P. (2022). *Overview of the WANLP 2022 shared task on propaganda detection in Arabic*. arXiv:*2211*.10057.

Albert, C. D., Mullaney, S., Huitt, J., Hunter, L., & Snider, L. (2023). Weaponizing words: Using technology to proliferate information warfare. *Cyber Defense Review*, *8*(3), 15–31..

Alhuzali, H., & Ananiadou, S. (2021). *SpanEmo: Casting multi-label emotion classification as span-prediction*. arXiv:*2101*.10038.

Ali, S. R., & Fahmy, S. (2013). Gatekeeping and citizen journalism: The use of social media during the recent uprisings in Iran, Egypt, and Libya. *Media, War & Conflict*, *6*(1), 55–69.

Alizadeh, M., Shapiro, J. N., Buntain, C., & Tucker, J. A. (2020). Content-based features predict social media influence operations. *Science Advances*, *6*(30), Article eabb5824.

Arif, A., Stewart, L. G., & Starbird, K. (2018). Acting the part: Examining information operations within# BlackLivesMatter discourse. *Proceedings of the ACM on Human-Computer Interaction*, *2*, 1–27.

Badawy, A., Ferrara, E., & Lerman, K. (2018, August). Analyzing the digital traces of political manipulation: The 2016 Russian interference Twitter campaign. In *2018 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM)* (pp. 258–265). IEEE.

Barrie, C., & Siegel, A. (2021). Kingdom of trolls? Influence operations in the Saudi Twittersphere. *Journal of Quantitative Description: Digital Media*, *1*, 1173–1199.

Bastos, M., & Farkas, J. (2019). "Donald Trump is my president!" The Internet Research Agency Propaganda Machine. *Social Media + Society*, *5*(3), 1–13.

Beck, M. (2020). The aggravated struggle for regional power in the Middle East: American allies Saudi Arabia and Israel versus Iran. *Global Policy*, *11*(1), 84–92.

Beckman, J., & Nigatu, G. (2021). Do political factors influence U.S. crude oil imports? *International Journal of Energy Economics and Policy*, *11*(4), 288–297.

Benkler, Y., Faris, R., & Roberts, H. (2018). *Network propaganda: Manipulation, disinformation, and radicalization in American politics*. Oxford University Press.

Bergh, A. (2020). Understanding Influence Operations in social media. *Journal of Information Warfare*, *19*(4), 110–131.

Blane, J. T., Bellutta, D., & Carley, K. M. (2022). Social-Cyber maneuvers during the COVID-19 vaccine initial rollout: Content analysis of tweets. *Journal of Medical Internet Research*, *24*(3), Article e34040.

Blane, J. T., Ng, L. H. X., & Carley, K. M. (2023). Analyzing social-cyber maneuvers for spreading COVID-19 pro- and anti-vaccine information. In T. Ginossar, S. F. A. Shah, & D. Weiss (Eds.), *Vaccine communication online: Counteracting misinformation, rumors and lies* (pp. 57–80. Springer.

Bond, S. (2021). Just 12 People are Behind most Hoaxes on Social Media, Research Shows. NPR, https://www.npr.org/2021/05/13/996570855/disinformation-dozen-test-facebooks-twitters-ability-to-curb-vaccine-hoaxes. Accessed November 1, 2023.

Brangetto, P., & Veenendaal, M. A. (2016, May). Influence cyber operations: The use of cyberattacks in support of influence operations. In *2016 8th International Conference on Cyber Conflict (CyCon)* (pp. 113–126). IEEE.

Broniatowski, D. A., Kerchner, D., Farooq, F., Huang, X., Jamison, A. M., Dredze, M., Quinn, S. Q., & Ayers, J. W. (2022). Twitter and Facebook posts about COVID-19 are less likely to spread misinformation compared to other health topics. *PLOS ONE*, *17*(1), Article e0261768.

Caldarelli, G., De Nicola, R., Del Vigna, F., Petrocchi, M., & Saracco, F. (2020). The role of bot squads in the political propaganda on Twitter. *Communications Physics*, *3*(1), Article 81.

Callanan, J. (2009). *Covert action in the Cold War: US Policy, intelligence and CIA operations*. I.B. Tauris.

Carley, K. M. (2020). Social cybersecurity: An emerging science. *Computational and Mathematical Organization Theory*, *26*(4), 365–381.

Chaudhry, I. (2014). Arab Revolutions: Breaking fear|# hashtags for change: Can Twitter generate social progress in Saudi Arabia. *International Journal of Communication*, *8*, 943–961.

Chernobrov, D., & Briant, E. L. (2022). Competing propagandas: How the United States and Russia represent mutual propaganda activities. *Politics*, *42*(3), 393–409.

Ciftci, S., & Tezcür, G. M. (2016). Soft power, religion, and anti-Americanism in the Middle East. *Foreign Policy Analysis*, *12*(3), 374–394.

Cohen, D., & Bar'el, O. (2017, October). The use of cyberwarfare in influence operations. In *Yuval Ne'eman Workshop for Science, Technology and Security*. Tel-Aviv University.

Conrad, J. (2011). Interstate rivalry and terrorism: An unprobed link. *Journal of Conflict Resolution*, *55*(4), 529–555.

Cordey, S. (2019). *Cyber influence operations: An Overview and comparative analysis*. CSS Cyber Defense Reports. https://doi.org/10.3929/ethz-b-000382358

Danaditya, A., Ng, L. H. X., & Carley, K. M. (2022). From curious hashtags to polarized effect: Profiling coordinated actions in Indonesian twitter discourse. *Social Network Analysis and Mining*, *12*(1), Article 105.

Darwich, M. (2020). Escalation in failed military interventions: Saudi and Emirati quagmires in Yemen. *Global Policy*, *11*(1), 103–112. https://doi.org/10.1111/1758-5899.12781

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). *Bert: Pre-training of deep bidirectional transformers for language understanding*. arXiv preprint arXiv:1810.04805.

Dilley, L., Welna, W., & Foster, F. (2022). *QAnon propaganda on Twitter as information warfare: Influencers, networks, and narratives*. arXiv:*2207*.05118.

DiResta, R., Goldstein, J. A., & Grossman, S. (2021). Middle East influence operations: Observations across social media takedowns. *Digital Activism and Authoritarian Adaptation in the Middle East*, 91. https://pomeps.org/wp-content/uploads/2021/08/POMEPS_Studies_43_Web.pdf#page=92

Fahmy, S., Wanta, W., & Nisbet, E. C. (2012). Mediated public diplomacy: Satellite TV news in the Arab world and perception effects. *International Communication Gazette*, *74*(8), 728–749.

Ferrara, E. (2020). *What types of COVID-19 conspiracies are populated by Twitter bots?* arXiv preprint arXiv:2004.09531.

Gause, F. G. (2016). The Future of US-Saudi relations: The kingdom and the power. *Foreign Affairs*, *95*(4), 114–126.

Guarino, S., Trino, N., Celestini, A., Chessa, A., & Riotta, G. (2020). Characterizing networks of propaganda on twitter: A case study. *Applied Network Science*, *5*(1), 1–22.

Howard, P. N., Bolsover, G., Kollanyi, B., Bradshaw, S., & Neudert, L. M. (2017). *Junk news and bots during the U.S. Election: What were Michigan voters sharing over Twitter?* (Data Memo 2017.1). Project on Computational Propaganda. https://demtech.oii.ox.ac.uk/wp-content/uploads/sites/12/2017/03/What-Were-Michigan-Voters-Sharing-Over-Twitter-v2.pdf

Howard, P. N., & Kollanyi, B. (2016). *Bots, #strongerin, and #brexit: Computational propaganda during the UK-EU referendum*. arXiv:160606356.

Howard, P. N., Woolley, S., & Calo, R. (2018). Algorithms, bots, and political communication in the US 2016 election: The challenge of automated political communication for election law and administration. *Journal of Information Technology & Politics*, *15*(2), 81–93.

Jamal, A. A., Keohane, R. O., Romney, D., & Tingley, D. (2015). Anti-Americanism and anti-interventionism in Arabic Twitter discourses. *Perspectives on Politics*, *13*(1), 55–73.

Jowett, G. S., & O'Donnell, V. (2018). *Propaganda & persuasion*. Sage.

Kießling, B., Homburg, J., Drozdzynski, T., & Burkhardt, S. (2020). State propaganda on Twitter: How Iranian propaganda accounts have tried to influence the international discourse on Saudi Arabia. In C. Grimme, M. Preuss, F. W. Takes, & A. Waldherr (Eds.), *Disinformation in open online media: First multidisciplinary international symposium* (pp. 182–197). Springer.

Larson, E. V., Darilek, R. E., Gibran, D., Nichiporuk, B., Richardson, A., Schwartz, L. H., & Thurston, C. Q. (2009). *Foundations of effective influence operations: A framework for enhancing army capabilities* (Vol. 654). Rand Corporation.

Llewellyn, C., Cram, L., Favero, A., & Hill, R. L. (2018, May). Russian troll hunting in a Brexit Twitter archive. In *Proceedings of the 18th ACM/IEEE on joint conference on digital libraries* (pp. 361–362). ACM.

Long, D. E. (2019). *The United States and Saudi Arabia: Ambivalent allies*. Routledge.

Lukito, J. (2020). Coordinating a multi-platform disinformation campaign: Internet research agency activity on three U.S. social media platforms, 2015 to 2017. *Political Communication*, *37*(2), 238–255.

Mair, D. (2016). # Westgate: A case study–how Al-Shabaab used Twitter during an ongoing attack." *Studies in Conflict & Terrorism*, *40*(1), 24–43.

Maschmeyer, L., Abrahams, A., Pomerantsev, P., & Yermolenko, V. (2023). Donetsk don't tell–'hybrid war'in Ukraine and the limits of social media influence operations. *Journal of Information Technology & Politics*, 1–16. https://doi.org/10.1080/19331681.2023.2211969

McNeill, L. (2019). Assumed Identity: Writing and reading testimony through and as Anne Frank 1. In J. R. Resina (Ed.), *Inscribed identities* (pp. 61–74). Routledge.

Mendelson, S. E., & Gerber, T. P. (2008). Us and them: Anti-American views of the Putin generation. *Washington Quarterly*, *31*(2), 131–150.

Miller, D. T. (2019). Topics and emotions in Russian Twitter propaganda. *First Monday*, *24*(5). https://doi.org/10.5210/fm.v24i5.9638

Mitts, T. (2019). From isolation to radicalization: Anti-Muslim hostility and support for ISIS in the West. *American Political Science Review*, *113*(1), 173–194.

Mohaddes, K., & Pesaran, M. H. (2016). Country-specific oil supply shocks and the global economy: A counterfactual analysis. *Energy Economics*, *59*, 382–399.

Moriarty, B. (2015). Defeating ISIS on Twitter. *Technology Science*. https://techscience.org/a/2015092904/

Nakayama, B. (2022). Democracies and the Future of Offensive (Cyber-Enabled). *Information Operations*, *7*(3), 49–65.

Neudert, L. M. N. (2018) Germany: A cautionary tale. In S. C. Woolley & P. N. Howard (Eds.), *Computational propaganda: Political parties, politicians, and political manipulation on social media*. Oxford University Press.

Neumayer, E., & Plümper, T. (2011). Foreign terror on Americans. *Journal of Peace Research*, *48*(1), 3–17.

Ng, L. H. X., & Carley, K. M. (2023). Popping the hood on Chinese balloons: Examining the discourse between U.S. and China-geotagged accounts. *First Monday*, *28*(8). https://doi.org/10.5210/fm.v28i8.13159

Ng, L. H. X., Moffitt, J. D., & Carley, K. M. (2022). *Coordinated through a web of images: Analysis of image-based influence operations from China, Iran, Russia, and Venezuela*. arXiv:2206.03576.

Nip, J., & Sun, C. (2022). Public diplomacy, propaganda, or what? China's communication practices in the South China Sea dispute on Twitter. *Journal of Public Diplomacy*, *2*(1), 43–68. https://doi.org/10.23045/jpd.2022.2.1.4

Nogara, G., Vishnuprasad, P. S., Cardoso, F., Ayoub, O., Giordano, S., & Luceri, L. (2022, June). The disinformation dozen: An exploratory analysis of covid-19 disinformation proliferation on twitter. In *Proceedings of the 14th ACM Web Science Conference 2022* (pp. 348–358). ACM.

Paszke, A., Gross, S., Chintala, S., Chanan, G., & Yang, E. (2017). *Automatic differentiation in PyTorch* [Conference session]. 31st Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, United States.

Patrikarakos, D. (2017). *War in 140 characters: How social media is reshaping conflict in the twenty-first century*. Hachette UK.

Pierri, F., Luceri, L., Jindal, N., & Ferrara, E. (2023, April). Propaganda and misinformation on Facebook and Twitter during the Russian invasion of Ukraine. In *Proceedings of the 15Th ACM Web Science Conference, 2023* (pp. 65–74). ACM.

Prier, J. (2017). Commanding the trend: Social media as information warfare. *Strategic Studies Quarterly*, *11*(4), 50–85.

Safaya, A., Abdullatif, M., & Yuret, D. (2020). Kuisail at Semeval-2020 task 12: BERT-CNN for offensive speech identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation* (pp. 2054–2059). International Committee for Computational Linguistics.

Schünemann, W. J. (2022). A threat to democracies? An overview of theoretical approaches and empirical measurements for studying the effects of disinformation. In M. D. Cavelty & A. Wenger (Eds.), *Cyber security politics* (pp. 32–47). Routledge.

Shakil, K. A., Tabassum, K., Alqahtani, F. S., & Wani, M. A. (2021). Analyzing user digital emotions from a holy versus non-pilgrimage city in Saudi Arabia on twitter platform. *Applied Sciences*, *11*(15), 6846.

Simon, T., Goldberg, A., Aharonson-Daniel, L., Leykin, D., & Adini, B. (2014). Twitter in the cross fire—The use of social media in the Westgate Mall terror attack in Kenya. *PLOS ONE*, *9*(8), Article e104136.

Singer, P. W., & Brooking, E. T. (2018) *LikeWar: The weaponization of social media*. Eamon Dolan Books.

Starbird, K., Arif, A., & Wilson, T. (2019). Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations. *Proceedings of the ACM on Human-Computer Interaction*, *3*, 1–26.

Suchkov, M. (2016). Contemporary Russia-Saudi relations: Building a bridge of cooperation over the abyss of discrepancies. *Iran and the Caucasus*, *20*(2), 237–251.

Sullivan, R. (2014). Live-tweeting terror: A rhetorical analysis of@ HSMPress_ Twitter updates during the 2013 Nairobi hostage crisis. *Critical Studies on Terrorism*, *7*(3), 422–433.

Theohary, C. A. (2018). Information Warfare: Issues for Congress. *Congressional Research Service Report*.

Uyheng, J., Magelinski, T., Villa-Cox, R., Sowa, C., & Carley, K. M. (2020). Interoperable pipelines for social cyber-security: Assessing Twitter information operations during NATO Trident Juncture 2018. *Computational and Mathematical Organization Theory*, *26*, 465–483.

Van Dijk, T. A. (Ed.) (2011). *Discourse studies: A multidisciplinary introduction*. Sage.

Weimann, G. (2010). Terror on Facebook, Twitter, and YouTube. *The Brown Journal of World Affairs*, *16*(2), 45–54.

Weiss, J. C. (2013). Authoritarian signaling, mass audiences, and nationalist protest in China. *International Organization*, *67*(1), 1–35.

Whyte, C., Thrall, A. T., & Mazanec, B. M. (Eds). (2020). *Information warfare in the age of cyber conflict*. Routledge Press.

Wolf, H. (2015). Paper is patient': Tweets from the '# AnneFrank of Palestine. *Textual Practice*, *29*(7), 1355–1374.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., & Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations* (pp. 38–45).

Woolley, S. C. (2022). Digital propaganda: The power of influencers. *Journal of Democracy*, *33*(3), 115–129.

Yeh, C. K., Wu, W. C., Ko, W. J., & Wang, Y. C. F. (2017, February). Learning deep latent space for multi-label classification. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 31, No. 1, pp. 2838–2844). AAAI Press.

Author Biographies

Craig Douglas Albert, PhD, is professor and director of the Master of Arts in Intelligence and Security Studies at Augusta University. His main areas of research include strategic cybersecurity, artificial intelligence, influence operations and information warfare, and Ethnic Conflict, with a specific interest in Chechnya. He has published in *Cyber Defense Review*, *Journal of Cyber Policy, Politics and the Life Sciences*, *Intelligence and National Security*, *Defense & Security Analysis*, and *Global Society*. His work has been supported by NSF and Office of Naval Research.

Ahmed Aleroud, PhD, is an associate professor in the School of Computer and Cyber Sciences. His research work focuses on machine learning computational approaches and social approaches to combat security and privacy attacks, including social engineering attacks, computer network intrusions, disinformation, social media propaganda, state-linked social media attacks, online extremism, and re-identification of private information on individuals or devices. He is also working on adversarial attacks on security systems. His research has been published in journals such as *IEEE Transactions*, *ACM Digital Threats Research & Practice*, *Computers & Security*, *Information Security and Applications*, *Information Systems Frontiers*, *Knowledge and Information Systems*. His work has been supported by NSF, Office of Naval Research, European IP Networks (RIPE), and the Department of Energy.

Yufan Yang, PhD, is an assistant professor of political science at Augusta University. Her research interests include international security, political violence, political regimes, propaganda, information technology, statistical modeling, machine learning, and computational social science in general.

Abdullah Melhem is a graduate research assistant at Augusta University's School of Computer and Cyber Sciences. With a master's degree in data science from Jordan University of Science and Technology, Abd has a strong background in advanced data analytics. Previously, he worked as a data scientist at Jordan Design and Development Bureau (JODDB) in Jordan.

Josh Rutland is a graduate of Augusta University's Masters in Intelligence and Security Studies program. His research focuses on information warfare, cybersecurity, terrorism, and biosecurity. His work has appeared in journals such as *Politics and the Life Sciences*, *Politics & Policy*, *Behavioral Sciences of Terrorism and Political Aggression*, *Journal of Cyber Policy*, *The Cyber Defense Review*, *and PLOS Global Public Health.*