frontiers | Frontiers in Computer Science

TYPE Original Research
PUBLISHED 31 August 2022
DOI 10.3389/fcomp.2022.929515

OPEN ACCESS

EDITED BY
Siew H. Chan,
University of North Georgia,
United States

REVIEWED BY
Yingtong Dou,
University of Illinois at Chicago,
United States
Pailin Trongmateerut,
Thammasat University, Thailand

*CORRESPONDENCE
Mirela Silva
msilva1@ufl.edu

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Human-Media Interaction, a section of the journal Frontiers in Computer Science

RECEIVED 26 April 2022
ACCEPTED 11 July 2022
PUBLISHED 31 August 2022

CITATION
Shi H, Silva M, Giovanini L, Capecci D, Czech L, Fernandes J and Oliveira D (2022) Lumen: A machine learning framework to expose influence cues in texts. *Front. Comput. Sci.* 4:929515. doi: 10.3389/fcomp.2022.929515

Lumen: A machine learning framework to expose influence cues in texts

Hanyu Shi[1†], Mirela Silva[1*†], Luiz Giovanini[1], Daniel Capecci[1], Lauren Czech[1], Juliana Fernandes[2] and Daniela Oliveira[1]

[1]Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL, United States, [2]Department of Advertising, University of Florida, Gainesville, FL, United States

Phishing and disinformation are popular social engineering attacks with attackers invariably applying influence cues in texts to make them more appealing to users. We introduce Lumen, a learning-based framework that exposes influence cues in text: (i) persuasion, (ii) framing, (iii) emotion, (iv) objectivity/subjectivity, (v) guilt/blame, and (vi) use of emphasis. Lumen was trained with a newly developed dataset of 3K texts comprised of disinformation, phishing, hyperpartisan news, and mainstream news. Evaluation of Lumen in comparison to other learning models showed that Lumen and LSTM presented the best F1-micro score, but Lumen yielded better interpretability. Our results highlight the promise of ML to expose influence cues in text, toward the goal of application in automatic labeling tools to improve the accuracy of human-based detection and reduce the likelihood of users falling for deceptive online content.

KEYWORDS

media analysis, deception, disinformation, misinformation, dataset

The Web has increasingly become an ecosystem for deception. Beyond social engineering attacks such as phishing which put Internet users and even national security at great peril (Mueller, 2019; 2021), false information is greatly shaping the political, social, and economic landscapes of our society, exacerbated and brought to light in recent years by social media. Recent years have undoubtedly brought to light the dangers of selective exposure[1], and false content can increase individuals' beliefs in the falsehood (Ross et al., 2021). These deceptive and divisive misuses of online media have evolved the previously seemingly tacit political lines to the forefront of our very own individual identities (Kalsnes and Larsson, 2021), thus raising concern for the anti-democracy effects caused by this polarization of our society (Barnidge and Peacock, 2019).

A key invariant of deceptive content is the application of influence cues in the text. Research on deception detection (Kahneman and Tversky, 1979; Russell, 1980; Cialdini, 1993; Rothman and Salovey, 1997; Kircanski et al., 2018) reveals that deceivers apply influence cues in messages to increase their appeal to the recipients. We posit several

1 A theory akin to *confirmation bias* and often used in Communication research pertaining to the idea that individuals favor information that reinforces their prior beliefs (Stroud, 2014).

types of influence cues that are relevant and prevalent in deceptive texts: (i) the principle of persuasion applied (Cialdini, 1993, 2001) (e.g., *authority*, *scarcity*), (ii) the framing of the message as either potentially causing a *gain* or a *loss* (Kahneman and Tversky, 1979; Rothman and Salovey, 1997), (iii) the positive/negative emotional salience/valence of the content (Russell, 1980; Kircanski et al., 2018), (iv) the subjectivity or objectivity of sentences in the text, (v) attribution of blame/guilt, and (vi) the use of emphasis.

Additionally, works such as Ross et al. (2021) found that the ability to think deliberately and analytically (i.e., "System 2", Kahneman and Tversky, 1979) is generally associated with the rejection of disinformation, regardless of the participants' political alignment—thus, the activation of this analytical thinking mode may act as an "antidote" to today's selective exposure. **We therefore advocate that interventions should mitigate deceptive content *via* the exposure of influence cues in texts.** Similar to the government and state-affiliated media account labels on Twitter Twitter Help Center, bringing awareness to the influence cues present in misleading texts may, in turn, aid users by providing additional context in the message, thus helping users think analytically, and benefit future work aimed at the automatic detection of deceptive online content.

Toward this goal, we introduce Lumen[2], a two-layer learning framework that exposes influence cues in text using a novel combination of well-known existing methods: (i) topic modeling to extract structural features in text; (ii) sentiment analysis to extract emotional salience; (iii) LIWC[3] to extract dictionary features related to influence cues; and (iv) a classification model to leverage the extracted features to predict the presence of influence cues. To evaluate Lumen's effectiveness, we leveraged our dataset[4] of 2,771 diverse pieces of online texts, manually labeled by our research team according to the influence cues in the text using standard qualitative analysis methods. We must, however, emphasize that Lumen is *not* a consumer-focused end-product, and instead is insomuch as a module for application in future user tools that we shall make publicly available to be leveraged by researchers in future work (as described in Section 5.2).

Our newly developed dataset is comprised of nearly 3K texts, where 1K were mainstream news articles, and 2K deceptive or misleading content in the form of: Russia's Internet Research Agency's (IRA) propaganda targeting Americans in the 2016 U.S. Presidential Election, phishing emails, and fake and hyperpartisan news articles. Here, we briefly define these terms, which we argue fall within the same "deceptive text umbrella." *Disinformation* constitutes any *purposefully* deceptive content aimed at altering the opinion of or confusing an individual or group. Within disinformation, we find instances of *propaganda* [facts, rumors, half-truths, or lies disseminated manipulatively for the purpose of influencing public opinion (Smith, 2021)] and *fake news* [fabricated information that mimics real online news (Ross et al., 2021) and considerably overlaps with hyperpartisan news (Barnidge and Peacock, 2019)]. *Mis*information's subtler, political form is *hyperpartisan news*, which entails a misleading coverage of factual events through the lens of a strong partisan bias, typically challenging mainstream narratives (Barnidge and Peacock, 2019; Ross et al., 2021). *Phishing* is a social engineering attack aimed at influencing users *via* deceptive arguments into an action (e.g., clicking on a malicious link) that will go against the user's best interests. Though phishing differs from disinformation in its modus operandi, we argue that it overlaps with misleading media in their main purpose—to galvanize users into clicking a link or button by triggering the victim's emotions (Barnidge and Peacock, 2019), and leveraging influence and deception.

We conducted a quantitative analysis of the dataset, which showed that *authority* and *commitment* were the most common principles of persuasion in the dataset (71% and 52%, respectively), the latter of which was especially common in news articles. Phishing emails had the largest occurrence of *scarcity* (65%). Framing was a relatively rare occurrence (13% *gain* and 7% *loss*), though *gain* framing was predominantly prevalent in phishing emails (41%). The dataset invoked an overall positive sentiment (VADER compound score of 0.232), with phishing emails containing the most positive average sentiment (0.635) and fake news with the most negative average sentiment (−0.163). *Objectivity* and *subjectivity* occurred in over half of the dataset, with *objectivity* most prevalent in fake news articles (72%) and *subjectivity* most common in IRA ads (77%). Attribution of blame/guilt was disproportionately frequent for fake and hyperpartisan news (between 38 and 45%). The use of emphasis was much more common in informal texts (e.g., IRA social media ads, 70%), and less common in news articles (e.g., mainstream media, 17%).

We evaluated Lumen in comparison with other traditional ML and deep learning algorithms. Lumen presented the best performance in terms of its *F*1-micro score (69.23%), performing similarly to LSTM (69.48%). In terms of *F*1-macro, LSTM (64.20%) performed better than Lumen (58.30%); however, Lumen presented better interpretability for intuitively understanding of the model, as it provides both the relative importance of each feature and the topic structure of the training dataset without additional computational costs, which cannot be obtained with LSTM as it operates as a black-box. Our results highlight the promise of exposing influence cues in text *via* learning methods.

This paper is organized as follows. Section 1 positions this paper's contributions in comparison to related work in the field.

2 LIWC (Pennebaker et al., 2015) is a transparent text analysis program that counts words in psychologically meaningful categories and is widely used to quantify psychometric characteristics of raw text data.

3 From Latin, meaning "to illuminate".

4 Available at: https://github.com/danielaoliveira/Potentiam.

Section 2 details the methodology used to generate our coded dataset. Section 3 describes Lumen's design and implementation, as well as Lumen's experimental evaluation. Section 4 contains a quantitative analysis of our dataset, and Lumen's evaluation and performance. Section 5 summarizes our findings and discusses the limitations of our work, as well as recommendations for future work. Section 6 concludes the paper.

1. Related work

This section briefly summarizes the extensive body of work on machine learning methods to automatically detect disinformation and hyperpartisan news, and initial efforts to detect the presence of influence cues in text.

1.1. Automatic detection of deceptive text

1.1.1. Phishing and spam

Most anti-phishing research has focused on automatic detection of malicious messages and URLs before they reach a user's inbox *via* a combination of blocklists (Dong et al., 2015; Oest et al., 2019) and ML (Peng et al., 2018; Bursztein and Oliveira, 2019). Despite yielding high filtering rates in practice, these approaches cannot prevent zero-day phishing[5] from reaching users because determining maliciousness of text is an open problem and phishing constantly changes, rendering learning models and blocklists outdated in a short period of time (Bursztein and Oliveira, 2019). Unless the same message has been previously reported to an email provider as malicious by a user or the provider has the embedded URL in its blocklist, determining maliciousness is extremely challenging. Furthermore, the traditional approach to automatically detect phishing takes a binary standpoint (phishing or legitimate, e.g., Chandrasekaran et al., 2006; Basnet et al., 2008; Shyni et al., 2016), potentially overlooking distinctive nuances and the sheer diversity of malicious messages.

Given the limitations of automated detection in handling zero-day phishing, human detection has been proposed as a complementary strategy. The goal is to either warn users about issues with security indicators in web sites, which could be landing pages of malicious URLs (Sunshine et al., 2009; Felt et al., 2015) or train users into recognizing malicious content online (Sheng et al., 2007). These approaches are not without their own limitations. For example, research on the effectiveness of SSL warnings shows that users either habituate or tend to ignore warnings due to false positives or a lack of understanding about the warning message (Akhawe and Felt, 2013; Vance et al., 2017).

5 A new, not-yet reported phishing email.

1.1.2. Fake and hyperpartisan media

The previously known "antidote" to reduce polarization and increase readers' tolerance to selective exposure was *via* the use of counter-dispositional information (Barnidge and Peacock, 2019). However, countering misleading texts with mainstream or high-quality content in the age of rapid-fire social media comes with logistical and nuanced difficulties. Pennycook and Rand (2021) provide a thorough review of the three main approaches employed in fighting misinformation: automatic detection, debunking by field experts (which is not scalable), and exposing the publisher of the news source.

Similar to zero-day phishing, disinformation is constantly morphing, such that "zero-day" disinformation may thwart already-established algorithms, such was the case with the COVID-19 pandemic (Pennycook and Rand, 2021). Additionally, the final determination of a *fake, true*, or *hyperpartisan* label is fraught with subjectivity. Even fact-checkers are not immune—their agreement rates plummet for ambiguous statements (Lim, 2018), calling into question their efficacy in hyperpartisan news.

We posit that one facet of the solution lies within the combination of human and automated detection. Pennycook and Rand (2021) conclude that lack of careful reasoning and domain knowledge is linked to poor truth discernment, suggesting (alongside Bago et al., 2020; Ross et al., 2021) that future work should aim to trigger users to think slowly and analytically (Kahneman and Tversky, 1979) while assessing the accuracy of the information presented. Lumen aims to fulfill the first step of this goal, as our framework exposes influence cues in texts, which we hypothesize is disproportionately leveraged in deceptive content.

1.2. Detecting influence in deceptive texts

1.2.1. Phishing

We focus on prior work that has investigated the extent to which Cialdini's principles of persuasion (PoP) (Cialdini, 1993, 2001) (described in Section 2) are used in phishing emails (Stajano and Wilson, 2011; Ferreira and Teles, 2019; Oliveira et al., 2019) and how users are susceptible to them (Lawson et al., 2017; Oliveira et al., 2017).

Lawson et al. (2017) leveraged a personality inventory and an email identification task to investigate the relationship between personality and Cialdini's PoP. The authors found that *extroversion* was significantly correlated with increased susceptibility to *commitment, liking,* and the pair *(authority, commitment)*, the latter of which was found in 41% of our dataset. Following Cialdini's PoP, after manually labeling ∼200 phishing emails, Akbar (2014) found that *authority* was the most frequent principle in the phishing emails, followed by *scarcity*,

corroborating our findings for high prevalence of *authority*. However, in a large-scale phishing email study with more than 2,000 participants, Wright et al. (2014) found that *liking* receives the highest phishing response rate, while *authority* received the lowest. Oliveira et al. (2017); Lin et al. (2019) unraveled the complicated relationship between PoP and Internet user age and susceptibility, finding that young users are most vulnerable to *scarcity*, while older ones are most likely to fall for *reciprocation*, with *authority* highly effective for both age groups. These results are promising in highlighting the potential usability of exposing influence cues to users.

1.2.2. Fake and hyperpartisan news

Contrary to phishing, few studies (e.g., Zhou and Zafarani, 2020, p. 76) have focused on detecting influence cues or analyzing how users are susceptible to them in the context of fake or highly partisan content. Xu et al. (2020) stands out as the authors used a mixed-methods analysis, leveraging both manual analysis of the textual content of 1.2K immigration-related news articles from 17 different news outlets, and computational linguistics (including, as we did, LIWC). The authors found that moral frames that emphasize that support *authority/respect* were shared/liked more, while the opposite occurred for *reciprocity/fairness*. Whereas we solely used trained coders, they measured the aforementioned frames by applying the moral foundations dictionary (Graham et al., 2009).

To the best of our knowledge, no prior work has investigated or attempted to automatically detect influence cues in texts in such a large dataset, containing multiple types of deceptive texts. In this work, we go beyond Cialdini's principles to also detect gain and loss framing, emotional salience, subjectivity and objectivity, and the use of emphasis and blame. Further, no prior work has made available to the research community a dataset of deceptive texts labeled according to the influence cues applied in the text.

2. Dataset curation and coding methodology

This section describes the methodology to generate the labeled dataset of online texts used to train Lumen, including the definition of each of the influence cues labels.

2.1. Curating the dataset

We composed a diverse dataset by gathering different types of texts from multiple sources, split into three groups (Table 1): **deceptive texts** (1,082 pieces of text containing disinformation and/or deception tactics), **hyperpartisan news** (1,003 hyperpartisan media news from politically right- and left-leaning publications), and **mainstream news** (974 center mainstream media news). Our dataset therefore contained 3,059 pieces of text in total.

TABLE 1 Description of curated dataset.

	Document type	Number of documents
Deceptive texts	Facebook ads	492
	Fake news	130
	Phishing emails	460
Hyperpartisan news	Right-leaning news	506
	Left-leaning news	497
Mainstream news	974	
Total	3,059	

For the **deceptive texts group**, we mixed 492 Facebook ads created by the Russian Internet Research Agency (IRA), 130 known fake news articles, and 460 phishing emails:

2.1.1. Facebook IRA ads

We leveraged a dataset of 3,517 Facebook ads created by the Russian IRA and made publicly available to the U.S. House of Representatives Permanent Select Committee on Intelligence (U.S. House of Representatives Permanent Selection Committee on Intelligence) by Facebook after internal audits. These ads were a small representative sample of over 80K organic content identified by the Committee and are estimated to have been exposed to over 126M Americans between June 2015 and August 2017. After discarding ads that did not have text entry, the dataset was decreased to 3,286 ads, which were mostly (52.8%) posted in 2016 (U.S. election year). We randomly selected 492 for inclusion.

2.1.2. Fake news

We leveraged a publicly available[6] dataset of nearly 17K news labeled as *fake* or *real* collected by (Sadeghi et al., 2020) from PolitiFact.com, a reputable source of fact-finding. We randomly selected 130 *fake* news ranging from 110 – 200 words dated between 2007 to 2020.

2.1.3. Phishing emails

To gather our dataset, we collected approximately 15K known phishing emails from multiple public sources (Smiles, 2019a,b,c,d,e,f,g,h). The emails were then cleaned and formatted to remove errors, noise (e.g., images, HTML elements), and any extraneous formatting so that only the raw email text remained.

6 https://ieee-dataport.org/open-access/fnid-fake-news-inference-dataset#files

We randomly selected 460 of these emails ranging from 50 − 150 words to be included as part of the Deceptive Texts.

For the **hyperpartisan news** and **mainstream news groups**, we used a public dataset[7] comprised of 2.7M news articles and essays from 27 American publications dated from 2013 to early 2020. We first selected articles ranging from 50 − 200 words and then classified them as *left*, *right*, or *center* news according to the AllSides Bias Rating[8]. For inclusion in the hyperpartisan news group, we randomly selected 506 *right* news and 497 *left* news; the former were dated from 2016 to 2017 and came from two publications sources (Breitbart and National Review) while the latter were dated from 2016 to 2019 and came from six publications (Buzzfeed News, Mashable, New Yorker, People, VICE, and Vox). To compose the mainstream news group, we randomly selected 974 *center* news from all seven publications (Business Insider, CNBC, NPR, Reuters, TechCrunch, The Hill, and Wired) dated from 2014 to 2019.

2.1.4. Coding process

We then developed coding categories and a codebook based on Cialdiani's principles of influence (Cialdini, 1993), subjectivity/objectivity, and gain/loss framing (Kahneman and Tversky, 1980). These categories have been used in prior works (e.g., Oliveira et al., 2017, 2019, and were adapted for the purposes of this study, as well as with the additional emphasis and blame/guilt attribution categories. Next, we held an initial training session with nine undergraduate students. The training involved a thorough description of the coding categories, their definitions and operationalizations, as well as a workshop-style training where coders labeled a small sample of the texts to get acquainted with the coding platform, the codebook, and the texts. Coders were instructed to read the text at least twice before starting the coding to ensure they understood it. After that, coders were asked to share their experiences labeling the texts and to discuss any issues or questions about the process. After this training session, two intercoder reliability pretests were conducted; in the first pretest, coders independently co-coded a sample of 20 texts, and in the second pretest, coders independently co-coded a sample of 40 texts. After each one of these pretests, a discussion and new training session followed to clarify any issues with the categories and codebook.

Following these additional discussion and training sessions, coders were then instructed to co-code 260 texts which served as our intercoder reliability sample. To calculate intercoder reliability, we used three indexes. *Cohen's kappa* and *Percent of Agreement* ranged from 0.40 to 0.90, and 66% to 99%, respectively, which was considered moderately satisfactory. Due to the nature of the coding and type of texts, we also opted to use *Perrault and Leigh's index* because (a) it has been used in similar studies that also use nominal data (Hove et al., 2013; Fuller and Rice, 2014; Morey and Eveland, 2016; Rice et al., 2018); (b) it is the most appropriate reliability measure for 0/1 coding (i.e., when coders mark for absence or presence of given categories), as traditional approaches do not take into consideration two zeros as agreement and thus penalize reliability even if coders disagree only a few times (Perreault and Leigh, 1989); and (c) indexes such as Cohen's kappa and Scott's pi have been criticized for being overly conservative and difficult to compare to other indexes of reliability (Lombard et al., 2002). Perrault and Leigh's index (I_r) returned a range of 0.67 to 0.99, which was considered satisfactory. Finally, the remaining texts were divided equally between all coders, who coded all the texts independently using an electronic coding sheet in Qualtrics. Coders were instructed to distribute their workload equally over the coding period to counteract possible fatigue effects. This coding process lasted 3 months.

2.1.5. Influence cues definitions

The coding categories were divided into five main concepts: *principles of influence*, *gain/loss framing*, *objectivity/subjectivity*, *attribution of guilt*, and *emphasis*. Coders marked for the absence (0) or presence (1) of each of the categories. Definitions and examples for each influence are detailed in Appendix 1 (Supplementary material), leveraged from the coding manual we curated to train our group of coders.

Principles of persuasion (PoP). *Persuasion* refers to a set of principles that influence how people concede or comply with requests. The principles of influence were based on Cialdini's marketing research work (Cialdini, 1993, 2001), and consist of the following six principles: (i) *authority*[9] or expertise, (ii) *reciprocation*, (iii) *commitment* and consistency, (iv) *liking*, (v) *scarcity*, and (vi) *social proof*. We added subcategories to the principles of commitment (i.e., indignation and call to action) and social proof (i.e., admonition) because an initial perusal of texts revealed consistent usage across texts.

Framing. Framing refers to the presentation of a message (e.g., health message, financial options, and advertisement) as implying a possible gain (i.e., possible benefits of performing the action) vs. implying a possible loss (i.e., costs of not performing a behavior) (Kahneman and Tversky, 1979; Rothman and Salovey, 1997; Kühberger, 1998). Framing can affect decision-making and behavior; work by Kahneman and Tversky (1979) on loss aversion supports the human tendency to prefer avoiding losses over acquiring equivalent gains.

Slant. Slant refers to whether a text is written subjectively or objectively; *subjective*sentences generally refer to a personal opinion/judgment or emotion, whereas *objective* sentences fired

7 https://components.one/datasets/all-the-news-2-news-articles-dataset/

8 https://www.allsides.com/media-bias/media-bias-ratings

9 e.g., people tend to comply with requests or accept arguments made by figures of authority.

to factual information that is based on evidence, or when evidence is presented. It is important to note that we did not ask our coders to fact check, instead asking them to rely on sentence structure, grammar, and semantics to determine the label of *objective* or *subjective*.

Attribution of blame/guilt. Blame or guilt refers to when a text references "another" person/object/idea for wrong or bad things that have happened.

Emphasis. Emphasis refers to the use of all caps text, exclamation points (either one or multiple), several question marks, bold text, italics text, or anything that is used to call attention in text.

3. Lumen design and implementation

This section describes the design, implementation, and evaluation of Lumen, our proposed two-level hierarchical learning-based framework to expose influence cues in texts.

3.1. Lumen overview

Exposing presence of persuasion and framing is tackled as a multi-labeling document classification problem, where zero, one, or more labels can be assigned to each document. Due to recent developments in natural language processing, emotional salience is an input feature that Lumen exposes leveraging sentiment analysis. Note that Lumen's goal is not to distinguish deceptive vs. benign texts, but to expose different influence cues applied in different types of texts.

Figure 1 illustrates Lumen's two-level hierarchical learning-based architecture. On the first level, the following features are extracted from the raw text: (i) topical structure inferred by topic modeling, (ii) LIWC features related to influence keywords (Pennebaker et al., 2015), and (iii) emotional salience features learned *via* sentiment analysis (Hutto and Gilbert, 2014). On the second level, a classification model is used to identify the influence cues existing in the text.

3.2. Topic structure features

Probabilistic topic modeling algorithms are often used to infer the topic structure of unstructured text data (Steyvers and Griffiths, 2007; Blei and Lafferty, 2009), which in our case are deceptive texts, hyperpartisan news, and mainstream news. Generally, these algorithms assume that a collection of documents (i.e., a corpus) are created following the generative process.

Suppose that there are D documents in the corpus $\mathcal{C}$ and each document $d = 1, ..., D$ has length m_d. Also suppose that there are in total K different topics in the corpus and the vocabulary $\mathcal{V}$ includes V unique words. The relations between documents and topics are determined by conditional probabilities $P(t|d)$, which specify the probability of topic $t = 1, ..., K$ given document d. The linkage between topics and unique words are established by conditional probabilities $P(w|t)$, which indicate the probability of word $w = 1, ..., V$ given topic t. According to the generative process, for each token $w(i_d)$, which denotes the i_d-th word in document d, we will first obtain the topic of this token, $z(i_d) = t$, according to $P(t|d)$. With the obtained $z(i_d)$, we then draw the a word $w(i_d) = w$ according to $P(w|t = z(i_d))$.

In this work, we leveraged Latent Dririchlet Allocation (LDA), one of the most widely used topic modeling algorithms, to infer topic structure in texts (Blei et al., 2003). In LDA, both $P(w|t)$ and $P(t|d)$ are assumed to have Dirichlet prior distributions. Given our dataset, which is the evidence to the probabilistic model, the goal of LDA is to infer the most likely conditional distribution $\hat{P}(w|t)$ and $\hat{P}(t|d)$, which is usually done by either variational Bayesian approach (Blei et al., 2003) or Gibbs Sampling (Griffiths and Steyvers, 2004). In Lumen, the conditional probabilities $\hat{P}(t|d)$ represent the topic structure of the dataset.

3.3. LIWC influence features

We use language to convey our thoughts, intentions, and emotions, with words serving as the basic building blocks of languages. Thus, the way different words are used in unstructured text data provide meaningful information to streamline our understanding of the use of influence cues in text data. Lumen thus leverages LIWC, a natural language processing framework that connects commonly-used words with categories (Tausczik and Pennebaker, 2010; Pennebaker et al., 2015) to retrieve influence features of texts to aid ML classification. LIWC includes more than 70 different categories in total, such as Perceptual Processes, Grammar, and Affect, and more than 6K common words.

However, not all the categories are related to influence. After careful inspection, we manually selected seven categories as features related to influence for Lumen. For persuasion, we selected the category *time* (related to *scarcity*); for emotion, we selected the categories *anxiety, anger,* and *sad*; and for framing, we selected the categories *reward* and *money* (gain), and *risk* (loss).

We denote the collection of the chosen LIWC categories as set $\mathcal{S}$. Given a text document d with document length m_d from the corpus $\mathcal{C}$, to build the LIWC feature $X_{i,d}^{LIWC}, \forall i \in \mathcal{S}$, we first count the number of words in the text d belonging to the LIWC category i, denoted as $n_{i,d}$, then normalize the raw word count with the document length:

$$X_{i,d}^{LIWC} = \frac{n_{i,d}}{m_d}, \forall i \in \mathcal{S}, d \in \mathcal{C}. \quad (1)$$

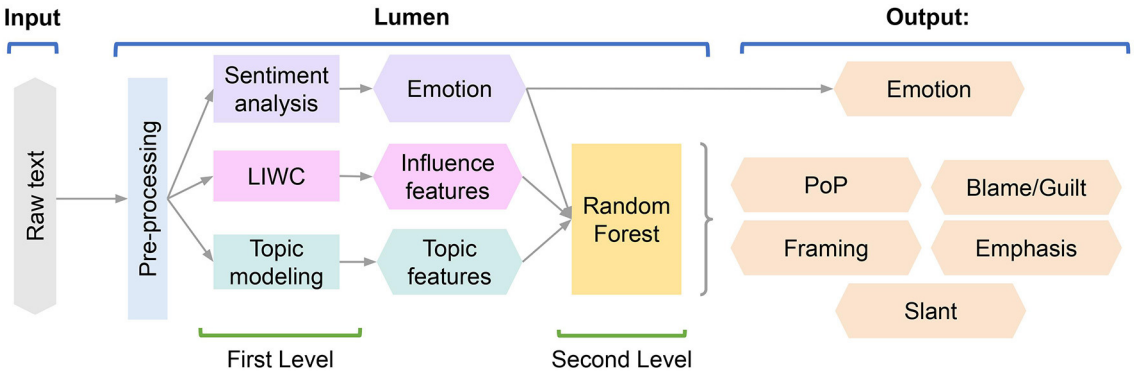

FIGURE 1
Lumen's two-level architecture. Pre-processed text undergoes sentiment analysis for extraction of emotional salience, LIWC analysis for extraction of features related to influence keywords, and topic modeling for structural features. These features are inputs to ML analysis for prediction of influence cues applied to the message.

3.4. Emotional salience using sentiment analysis

Emotional salience refers to both valence (positive to negative) and arousal (arousing to calming) of an experience or stimulus (Russell, 1980; Peace and Sinclair, 2012; Kircanski et al., 2018), and research has shown that deception detection is reduced for emotional compared to neutral stimuli (Peace and Sinclair, 2012). Similarly, persuasion messages that generate high (compared to low) arousal lead to poorer consumer decision-making (Kircanski et al., 2018). Emotional salience may impair full processing of deceptive content and high arousal content may trigger System 1, the fast, shortcut-based brain processing mode (Ariely et al., 2009).

In this work, we used a pre-trained rule-based model, VADER, to extract the emotional salience and valence from a document (Hutto and Gilbert, 2014). Both levels of emotion range from 0 to 1, where a small value means low emotional levels and a large number means high emotional levels. Therefore, emotional salience is both an input feature to the learning model and one of Lumen's outputs (see Figure 1).

3.5. Machine learning to predict persuasion and framing

Lumen's second level corresponds to the application of a general-purpose ML algorithm for document classification. Although Lumen is general enough to allow application of any general-purpose algorithm, in this paper, we applied Random Forest (RF) because it can provide the level of importance for each input predicative feature without additional computational cost, which aids in model understanding. Another advantage provided by RF is its robustness to the magnitudes of input predicative features, i.e., RF does not need feature normalization. We use the grid search approach to fine-tune the parameters in the RF model and follow the cross-validation to overcome any over-fitting issues of the model.

3.5.1. Dataset pre-processing

As described previously, Lumen generates three types of features at its first hierarchical level (emotional salience, LIWC categories, and topic structure), which serve as input for the learning-based prediction algorithm (Random Forest, for this analysis) at Lumen's second hierarchical level (Figure 1); these features rely on the unstructured texts in the dataset. However, different features need distinct preprocessing procedures. In our work, we used the Natural Language Toolkit (NLTK) (NLTK, 2020) to pre-process the dataset. For all three types of features, we first removed all the punctuation, special characters, digital numbers, and words with only one or two characters. Next, we tokenized each document into a list of lowercase words.

For topic modeling features, we removed stopwords (which provide little semantic information) and applied stemming (replacing a word's inflected form with its stem form) to further clean-up word tokens. For LIWC features, we matched each word in each text with the pre-determined word list in each LIWC category; we also performed stemming for LIWC features. We did not need to perform pre-processing for emotional salience because we applied NLTK (Hutto and Gilbert, 2014), which has its own tokenization and pre-processing procedures.

Additionally, we filtered out documents with less than ten words since topic modeling results for extremely short documents are not reliable (Shi et al., 2019). We were then left with 2,771 cleaned documents, with 183,442 tokens across the corpus, and 14,938 unique words in the vocabulary.

3.5.2. Training and testing

Next, we split the the 2,771 documents into a training and a testing set. In learning models, hyper-parameters are of crucial importance because they control the structure or learning processing of the algorithms. Lumen applies two learning algorithms: an unsupervised topic modeling algorithm, LDA, on the first hierarchical level and RF on the second level. Each algorithm introduces its own types of hyper-parameters; for LDA, examples include the number of topics and the concentration parameters for Dirichlet distributions, whereas for RF are the number of trees and the maximum depth of a tree. We also used the grid search approach to find a better combination of hyper-parameters. Note that due to time and computational power constraints, it is impossible to search all hyper-parameters and all their potential values. In this work, we only performed the grid search for number of topics (LDA) and the number of trees (RF). The results show that the optimal number of topics is 10 and the optimal number of trees in RF is 200. Note also that the optimal result is limited by the grid search space, which only contains a finite size of parameter combinations.

If we only trained and tested Lumen on one single pair of training and testing sets, there would be high risk of overfitting. To lower this risk, we used 5-fold cross-validation, wherein the final performance of the learning algorithm is the average performance over the five training and testing pairs.

In this work, we use the python "keras" package to generate the LSTM model. The LSTM model starts with a 50-dimensional "Embedding" layer followed with the bidirectional "LSTM" layer with 100 dimension output, and ends with a "Dense" layer with "sigmoid" activation function. The LSTM model is trained with 10 epochs. As for the split of the training/testing dataset, we are using the 5-fold cross validation to randomly split the dataset (80 Admittedly, there exists more advanced network-based algorithms, such as BERT, which may perform better than the algorithm selected in the paper. However, the aim of this paper is not to fully invest the performance of the network-based algorithms, LSTM is selected as a base line.

3.5.3. Evaluation metrics

To evaluate our results (see Section 4), we compared Lumen's performance in predicting the influence cues applied to a given document with three other document classification algorithms: (i) Labeled-LDA, (ii) LSTM, and (iii) naïve algorithm.

Labeled-LDA is a semi-supervised variation of the original LDA algorithm (Ramage et al., 2009; van der Heijden and Allodi, 2019). When training the Labeled-LDA, both the raw document and the human coded labels for influence cues were input into the model. Compared to Lumen, Labeled-LDA only uses the word frequency information from the raw text data and has a very rigid assumption of the relation between the word frequency information and the coded labels, which limits its flexibility and prediction ability.

TABLE 2 Evaluation metric results for different learning algorithms in detecting influence cues.

Algorithm	F1-macro (%)	F1-micro (%)	Overall accuracy (%)
Lumen	58.30	**69.23**	**72.43**
Labeled-LDA	52.35	60.55	64.22
LSTM	**64.20**	**69.48**	72.34
Naive	43.55	46.80	49.58

The bold values show that the prediction performance of our newly proposed algorithm is on par with that of other widely-used algorithms.

Long Short-Term Memory (LSTM) takes the input data recurrently, regulates the flow of information, and determines what to pass on to the next processing step and what to forget. Since neural networks mainly deal with vector operations, we used 50-dimensional word embedding matrix to transfer each word into vector space (Naili et al., 2017). The main shortcoming of neural network is that it works as a blackbox, making it difficult to understand the underlying mechanism.

The **naïve algorithm** served as a base line for our evaluation. We randomly generated each label for each document according to a Bernoulli distribution with equal probabilities for two outcomes.

As shown in Table 2, we used $F1$-score [following the work by Ramage et al. (2009) and van der Heijden and Allodi (2019)], and accuracy rate to quantify the performance of the algorithms. We note that the comparison of F-scores is only meaningful under the same experiment setup. It would be uninformative to compare F-scores from distinctive experiments in different pieces of work in the literature due to varying experiment conditions. $F1$-score can be easily calculated for single-labeling classification problems, where each document will only be assigned to one label. However, in our work, we are dealing with a multi-labeling classification problem, which means that no limit is imposed on how many labels each document can include. Thus, we employed two variations of the $F1$-score to quantify the overall performance of the learning algorithm: macro and micro $F1$-scores.

The $F1$-micro and $F1$-macro scores are two different approaches to quantify the performance of machine learning algorithms for multi-class prediction tasks. They are the variants of the same F1-score from two different aspect. The $F1$-macro weighs each class equally and $F2$-micro weighs each sample equally. In this paper, we use these two $F1$-scores to show that the prediction performance of our newly proposed algorithm is on par with that of other widely-used algorithms.

4. Results

This section details Lumen's evaluation. First, we provide a quantitative analysis of our newly developed dataset used to train

Lumen, and the results of Lumen classification in comparison to other ML algorithms.

4.1. Quantitative analysis of the dataset

We first begin by quantifying the curated dataset of 2,771 deceptive, hyperpartisan, or mainstream texts, hand-labeled by a group of coders. When considering all influence cues, most texts used between three and six cues per texts; only 3% of all texts leveraged a single influence cue, and 2% used zero cues ($n = 58$).

When considering the most common pairs and triplets between all influence cues, slant (i.e., *subjectivity* or *objectivity*) and principles of persuasion (PoP) dominated the top 10 most common pairings and triplets. As such, the most common pairs were *(authority, objectivity)* and *(authority, subjectivity)*, occurring for 48% and 45% of all texts, respectively. The most common (PoP, PoP) pairing was between *authority* and *commitment*, co-occurring in 41% of all texts. *Emphasis* appeared once in the top 10 pairs and twice in the top triplets: *(emphasis, subjectivity)* occurring for 29% of texts, and *(emphasis, authority, subjectivity)* and *(emphasis, commitment, subjectivity)* for 20% and 19% of texts, respectively. *Blame/guilt* appeared only once in the top triplets as *(authority, blame/guilt, objectivity)*, representing 19% of all texts. *Gain* framing appeared only as the 33rd most common pair *(gain, subjectivity)* and 18th most common triplet *(call to action, scarcity, gain)*, further emphasizing its scarcity in our dataset.

4.1.1. Principles of persuasion

We found that most texts in the dataset contained one to four principles of persuasion, with only 4% containing zero and 3% containing six or more PoP labels; 29% of texts apply two PoP and 23% leverage three PoP. Further, Figure 2 shows that *authority* and *commitment* were the most prevalent principles appearing, respectively, in 71% and 52% of the texts; meanwhile, *reciprocation* and *indignation* were the least common PoP (5% and 9%, respectively).

Almost all types of texts contained every PoP to varying degrees; the only exception is *reciprocation* (the least-used PoP overall) which was not at all present in fake news texts (in the Deceptive Texts Group) and barely present ($n = 3$, 0.6%) in right-leaning hyperpartisan news. *Authority* was the most-used PoP for all types of texts, except phishing emails (most: *call to action*) and IRA ads (most: *commitment*), both of which are in the Deceptive Texts Group.

Deceptive texts. Fake news was notably reliant on *authority* (92% of all fake news leveraged the *authority* label) compared to phishing emails (45%) and the IRA ads (32%); however, fake news used *liking, reciprocation*, and *scarcity* (5%, 0%, 3%, respectively) much less often than phishing emails (27%, 8%, 65%) or IRA ads (41%, 10%, 24%). Interestingly, *admonition* was most used by fake news (35%), though overall, *admonition* was only present in 14% of all texts. Phishing emails were noticeably more reliant on *call to action* (80%) and *scarcity* (65%) compared to fake news (33%, 3%) and IRA ads (40%, 24%), yet barely used *indignation* (0.4%) compared to the same (13% for fake news and 17% for IRA ads). The IRA ads relied on *indignation, liking, reciprocation*, and *social proof* much more than the others; note again that *reciprocation* was the least occurring PoP (5% overall), but was most commonly occurring in IRA ads (10%).

Hyperpartisan news. Right-leaning texts had nearly twice as much *call to action* and *indignation* than left-leaning texts (61% and 19% vs. 31% and 8%, respectively). Meanwhile, left-leaning hyperpartisan texts had noticeably more *liking* (30% vs. 13%), *reciprocation* (8% vs. 0.6%), and *scarcity* (27% vs. 13%) than right-leaning texts.

Mainstream news. *Authority* (88%) and *commitment* (43%) were the most frequently appearing PoP in center news, though this represents the third highest occurrence of *authority* and lowest use of *commitment* across all six text type groups. Mainstream news also used very little *indignation* (3%) compared the other text types except phishing emails (0.4%), and also demonstrated the lowest use of *social proof* (7%).

Authority and *commitment* were the most common PoP in the dataset, with the former most common in fake news articles. Phishing emails had the largest occurrence of *scarcity*.

4.1.2. Framing

There were few *gain* or *loss* labels for the overall dataset (only 13% and 7%, respectively). Very few texts (18%) were framed exclusively as either *gain* or *loss*, 81% did not include any framing at all, and only 1% of the texts used both *gain* and *loss* framing in the same message. We also found that *gain* was much more prevalent than *loss* across all types of texts, except for fake news, which showed an equal amount (1.5% for both *gain* and *loss*). Notably, phishing emails had significantly more *gain* and *loss* than any other text type (41% and 29%, respectively); mainstream center news and IRA ads showed some use of *gain* framing (10% and 13%, respectively) compared to the remaining text types.

Next, we investigated how persuasion and framing were used in texts by analyzing the pairs and triplets between the two influence cues. *Gain* framing most frequently occurred with *call to action* and *commitment*, though these represent only 9% of pairings. *Gain, call to action* and *scarcity* was the most common triplet between PoP and framing, occurring for 7% of all texts—this is notable as phishing emails had *call to action* and *scarcity* as its top PoP, and *gain* framing was also most prevalent in phishing. Also of note is that *loss* appeared in even fewer common pairs and triplets compared to *gain* (e.g., *loss* and *call to action* appeared in just 5% of texts).

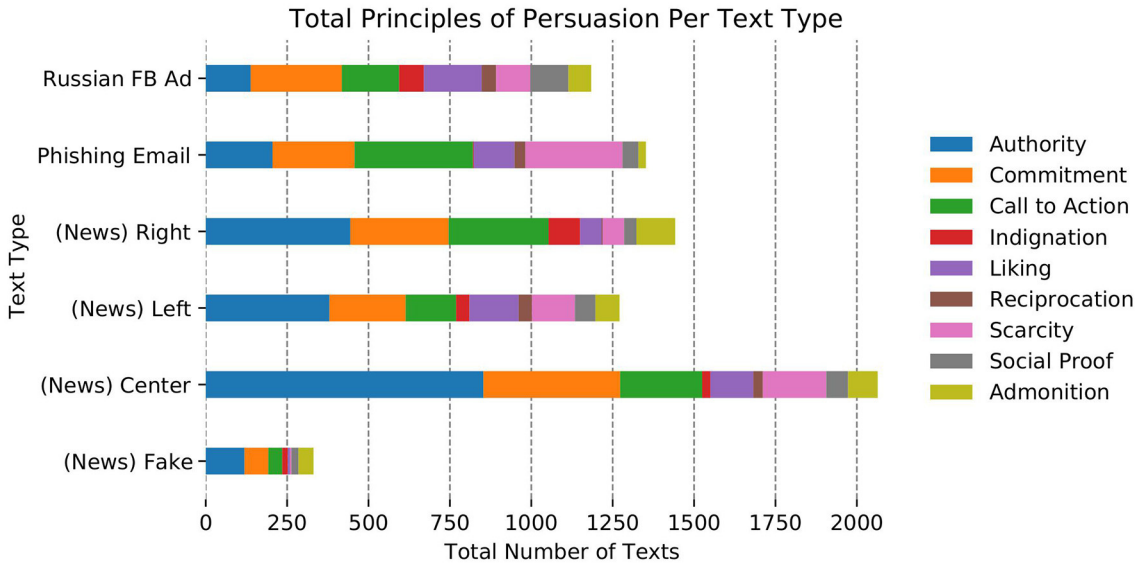

FIGURE 2
The total for each text type, broken down based on principles of persuasion.

Framing was a relatively rare occurrence in the dataset, though predominantly present in phishing emails, wherein *gain* was invoked 1.5× more often than *loss*.

4.1.3. Emotional salience

We used VADER's compound sentiment score (E, wherein $E \geq 0.05$, $E \leq -0.05$, and $-0.05 < E < 0.05$ denote positive, negative, and neutral sentiment, respectively) and LIWC's positive and negative emotion word count metrics to measure sentiment. Overall, our dataset was slightly positive in terms of average compound sentiment ($\mu = 0.23$) and with an average of 4.0 positive emotion words and 1.7 negative emotion words per text.

In terms of specific text types, fake news contained the only negative average compound sentiment (-0.163), and right-leaning hyperpartisan news had the only neutral average compound sentiment (0.015); all other text types had, on average, positive sentiment, with phishing emails as the most positive text type (0.635). Left-leaning hyperpartisan news had the highest average positive emotion word count (5.649) followed by phishing emails (4.732), whereas fake news had the highest average negative word count (2.892) followed by left hyperpartisan news (2.796).

We also analyzed whether emotional salience has indicative powers to predict the influence cues. Most influence cues and LIWC categories had an average positive sentiment, with *liking* and *gain* framing having the highest levels of positive emotion. *Anxiety* and *anger* (both LIWC categories) showed the only neutral sentiment, whereas *admonition, blame/guilt*, and *indignation* as the only negative sentiment (with the latter being the most negative out of all categories). Interestingly, items such as *loss* framing and LIWC's *risk* both had positive sentiment.

The dataset invoked an overall positive sentiment, with phishing emails containing the most positive average sentiment and fake news with the most negative average sentiment.

4.1.4. Slant

The *objective* and *subjective* labels were present in 52% and 64% of all texts in the dataset, respectively. This $>$ 50% frequency for both categories was present in all text types except phishing emails and IRA ads, where *subjectivity* was approximately 2.5× more common than *objectivity*. The most subjective text type were IRA ads (77%) and the most objective texts were fake news (72%); inversely, the least objective texts were phishing emails (27%) and least subjective were mainstream center news (58%).

More notably, there was an overlap between the slants, wherein 29% of all texts contained both *subjective* and *objective* labels. This could reflect mixing factual (objective) statements with subjective interpretations of them. Nonetheless, *objectivity* and *subjectivity* were independent variables, $\chi^2(4, N = 2,998) = 72.0, p \approx 0$. The parings *(objectivity, authority)* and *(subjectivity, authority)* were the top two most common pairs considering PoP and slant; these pairs occurred at nearly the same frequency within the dataset (48% and 45%, respectively). This pattern repeats itself for other (PoP, slant) pairings and triplets, insofar as *(objectivity, subjectivity, authority)* is the third most commonly occurring triplet. When comparing just (PoP, slant) triplets, slant is present in 9/10 top triplets, with *(subjectivity, authority, commitment)* and *(objectivity,*

authority, commitment) as the two most common triplets (30% and 27%, respectively).

Objectivity and *subjectivity* occurred over half of the dataset, with the latter was much more common in phishing emails and IRA ads, while the former was most common in fake news articles.

4.1.5. Attribution of blame and guilt

Twenty-nine percent of all texts contained the *blame/guilt* label. Interestingly, nearly the same proportions of fake news (45.4%) and right-leaning hyperpartisan news (45.0%) were labeled with *blame/guilt*, followed by left-leaning hyperpartisan news (38%). Phishing emails, IRA ads, and mainstream center media used *blame/guilt* at the lowest frequencies (ranging from 15 to 25%).

Blame/guilt was somewhat seen in the top 10 pairs with PoP, only pairing with *authority* (4th most common pairing with 26% frequency) and *commitment* (6th most common, 18%). However, *blame/guilt* appeared more frequently amongst the top 10 triplets with PoP, co-occurring with *authority, commitment, call to action,* and *social admonition*.

Blame/guilt was disproportionately frequent for fake and hyperpartisan news, commonly co-occurring with *authority* or *commitment*.

4.1.6. Emphasis

Emphasis was used in nearly 35% of all texts in the dataset. Among them, all news sources (fake, hyperpartisan, and mainstream) appeared with the smallest use of *emphasis* (range: 17% to 26%). This follows as news (regardless of veracity) likely is attempting to purport itself as legitimate. On the other hand, phishing emails and IRA ads were both shared on arguably more informal environments of communication (email and social media), and were thus often found to use emphasis (over 54% for both categories). Additionally, similar to previous analyses for other influence cues, *emphasis* largely co-occurred with *authority, commitment,* and *call to action*.

The use of emphasis was much more common in informal text types (phishing emails and IRA social media ads), and less common in news-like sources (fake, hyperpartisan, or mainstream).

4.1.7. LIWC features of influence

We also explored whether LIWC features have indicative powers to predict the influence cues. Supplementary Table 1 in Supplementary material shows that *indignation* and *admonition* had the highest average *anxiety* feature, while *liking* and *gain* framing had the lowest. *Indignation* also scored three times above the overall average for the *anger* feature, as well as for *sadness* (alongside *blame/guilt*), whereas *gain* had the lowest average for both *anger* and *sadness*. The *reward* feature was seen most in *liking* and in *gain*, while *risk* was slightly more common in *loss* framing. The *time* category had the highest overall average and was most common in *blame/guilt*, while *money* had the second largest overall average and was most common in *loss*.

We also saw that left-leaning hyperpartisan news had the highest average *anxiety*, *sadness*, *reward*, and *time* counts compared to all text types, whereas right-leaning hyperpartisan news averaged slightly higher than left-leaning media only in the *risk* feature. Note, however, that LIWC is calculated based on word counts and is therefore possibly biased toward longer length texts; it should thus be noted that while hyperpartisan left media had the highest averages for four of the seven LIWC features, hyperpartisan media also had the second largest average text length compared to other text types.

For the Deceptive Texts Group, phishing emails had the largest *risk* and *money* averages over all text types, while averaging lowest in *anxiety*, *anger*, and *sadness*. Fake news was highest overall in *anger*, though it was slightly higher in *anxiety*, *sadness*, and *time* compared to phishing emails and IRA ads. On the other hand, the IRA ads were lowest in *reward*, *risk*, *time*, and *money* compared to the its group.

Lastly, mainstream center media had no LIWC categories in either high or low extremities—most of its average LIWC values were close to the overall averages for the entire dataset.

LIWC influence features varied depending on the type of text. Left hyperpartisan news had the highest averages for four features (*anxiety*, *sadness*, *reward*, and *time*). Phishing evoked *risk* and *money*, while fake news evoked *anger*.

4.2. Lumen's multi-label prediction

This section describes our results in evaluating Lumen's multi-label prediction using the dataset. We compared Lumen's performance against three other ML algorithms: Labeled-LDA, LSTM, and a naïve algorithm. The former two learning algorithms and Lumen performed much better than the naïve algorithm, which shows that ML is promising for retrieval of influence cues in texts. From Tables 2, 3, we can see that Lumen's performance is as good as the state-of-the-art prediction algorithm LSTM in terms of $F1$-micro score and overall-accuracy (with $< 0.25\%$ difference between each metric). On the other hand, LSTM outperformed Lumen in terms of $F1$-macro, which is an unweighted mean of the metric for each labels, thus potentially indicating that Lumen underperforms LSTM in some labels although both algorithms share similar overall prediction result (accuracy). Nonetheless, Lumen presented better interpretability than LSTM (discussed below). Finally, both Lumen and LSTM presented better performance than Labeled-LDA in both $F1$-scores and accuracy, further emphasizing that additional features besides topic structures can help improve the performance of the prediction algorithm.

TABLE 3 Lumen's per-class performance compared to LSTM.

Influence cue	F1-Lumen (%)	F1-LSTM (%)
Authority	87.83	86.82
Commitment	63.61	61.46
Call to action	58.47	75.05
Subjectivity	78.68	74.07
Gain framing	30.89	46.83
Blame/guilt	47.87	57.05
Emphasis	41.55	53.74

TABLE 4 The top-five most important features for Lumen's prediction.

Input feature	Importance	Keywords
Topic-1	0.073	*account, bank, security, time*
Positive sentiment	0.071	N/A
Negative sentiment	0.070	N/A
Topic-8	0.065	*report, share, billion, source, profit*
Topic-2	0.062	*black, people, trump, police, twitter*

Topic features are related to LDA topic modeling results.

To show Lumen's ability to provide better understanding to practitioners (i.e., interpretability), we trained it with our dataset and the optimal hyper-parameter values from grid search. After training, Lumen provided both the relative importance of each input feature and the topic structure of the dataset without additional computational costs, which LSTM cannot provide because it operates as a black-box.

Table 4 shows the top-five important features in Lumen's prediction decision-making process. Among these features, two are related to sentiment, and the remaining three are topic features (related to bank account security, company profit report, and current events tweets), which shows the validity for the choice of these types of input features. Positive and negative sentiment had comparable levels of importance to Lumen, alongside the bank account security topic.

Here, we add a case study of phishing email to briefly show how the Lumen framework works:

> Date: 30th-April-2015
> Greetings from Skrill.co.uk Please take the time to read this message - it contains important information about your Skrill account. This is an urgent reminder for you as your account has been flagged by our Skrill team. Please login now to confirm your details or we will have no choice but to suspend your Skrill account. Click on the link below to log in your online banking for verification: http://account.skrill.com.login.locale-enUK.twehea.skrlnewdd...

In the Lumen framework, we first use a pre-trained rule-based model, VADER, to extract the emotional salience. The result shows that this document scores low both in positive (0.091) and negative sentiments (0.052), and high on neutral sentiment (0.857). Note that the sum of positive, negative, and neutral sentiment scores is 1. Then, we use the LIWC dataset to obtain the LIWC influence features. For example, in the above phishing email there are many words from the LIWC *money* category (e.g., "account," "banking," etc.) and *time* category (e.g., "time," "now," etc.). The text also has a LIWC *money* feature score of 11.9% and LIWC *time* feature score of 9.5%. Then, the phishing email text is input to the trained topic model to extract the embedded topic structure vector. Finally, the sentiment scores, the LIWC influence features, and the topic structure vector will be inserted to the trained Random Forest model to predict the persuasion and framing. In this specific case, the trained Lumen framework successfully predicted that there are "Authority," "Commitment," "Subjectivity," and "Loss framing" in the document.

5. Discussion

In this paper, we posit that interventions to aid human-based detection of deceptive texts should leverage a key invariant of these attacks: the application of influence cues in the text to increase its appeal to users. The exposure of these influence cues to users can potentially improve their decision-making by triggering their analytical thinking when confronted with suspicious texts, which were not flagged as malicious *via* automatic detection methods. Stepping toward this goal, we introduced Lumen, a learning framework that combines topic modeling, LIWC, sentiment analysis, and ML to expose the following influence cues in deceptive texts: persuasion, gain or loss framing, emotional salience, subjectivity or objectivity, and use of emphasis or attribution of guilt. Lumen was trained and tested on a newly developed dataset of 2,771 texts, comprised of purposefully deceptive texts, and hyperpartisan and mainstream news, all labeled according to influence cues.

5.1. Key findings

Most texts in the dataset applied between three and six influence cues; we hypothesize that these findings may reflect the potential appeal or popularity of texts of moderate complexity. Deceptive or misleading texts constructed without any influence cues are too simple to convince the reader, while texts with too

many influence cues might be far too long or complex, which are in turn more time-consuming to write (for attackers) and to read (for receivers).

Most texts also applied *authority*, which is concerning as it has been shown to be one of the most impactful in user susceptibility to phishing studies (Oliveira et al., 2017). Meanwhile, *reciprocation* was the least used principle at only 5%; this may be an indication that *reciprocation* does not lend itself well to be applied in text, as it requires giving something to the recipient first and expecting an action in return later. Nonetheless, *reciprocation* was most common in IRA ads (10%); these ads were posted on Facebook, and social media might be a more natural and intuitive location to give gifts or compliments. We also found that the application of the PoP was highly imbalanced with *reciprocation, indignation, social proof,* and *admonition* each being applied less than 15% the texts during the coding process.

The least used influence cue were *gain* and *loss* framing, appearing in only 13% and 7% of all texts. Though Kahneman and Tversky (1979) posited that *loss* is more impactful than the possibility of a gain, our dataset indicates that *gain* was more prevalent than *loss*. This is especially the case in phishing emails, wherein the framing frequencies increase to 41% and 29%; this difference suggests that in phishing emails, attackers might be attempting to lure users to potential financial gain. We further hypothesize that phishing emails exhibited these high rates of framing because successful phishing survives only *via* a direct action from the user (e.g., clicking a link), which may therefore motivate attackers to implement framing as a key influence method. Phishing emails also exhibited the most positive average sentiment (0.635) compared to other text types, possibly related to its large volume of *gain* labels, which were also strongly positive in sentiment (0.568).

Interestingly, texts varied among themselves in terms of influence cues even within their own groups. For example, within the Deceptive Texts Group, fake news used notably more *authority*, *objectivity*, and *blame/guilt* compared to phishing emails and IRA ads, and was much lower in sentiment compared to the latter two. Though phishing emails and IRA ads were more similar, phishing was nonetheless different in its use of higher positive sentiment, *gain* framing, *scarcity*, and lower *blame/guilt*. This was also evident within the Hyperpartisan News Group—while right-learning news had a higher frequency of *commitment*, *call to action*, and *admonition* than left-learning news, the opposite was also true for *liking*, *reciprocation*, and *scarcity*. Even comparing among all news types (fake, hyperpartisan, and mainstream), this diversity of influence cues still prevailed, with the only resounding agreement in a relative lack of use of *emphasis*. This diversity across text types gives evidence of the highly imbalanced application of influence cues in real deceptive or misleading campaigns.

We envision the use of Lumen (and ML methods in general) to expose influence cues as a promising direction for application tools to aid human detection of cyber-social engineering and disinformation. Lumen presented a comparable performance compared to LSTM in terms of the $F1$-micro. Lumen's interpretability can allow a better understanding of both the dataset and the decision-making process Lumen undergoes, consequently providing invaluable insights for feature selection.

5.2. Limitations and future work

5.2.1. Dataset

One of the limitations of our work is that the dataset is unbalanced. For example, our coding process revealed that some influence (e.g., *authority*) were disproportionately more prevalent than others (e.g., *reciprocation*, framing). Even though an unbalanced dataset is not ideal for ML analyses, we see this as part of the phenomenon. Attackers and writers might find it more difficult to construct certain concept *via* text, thus favoring other more effective and direct influence cues such as *authority*. Ultimately, our dataset is novel in that each of the nearly 3K items were coded according to 12 different variables; this was a time-expensive process and we shall test the scalability of Lumen in future work. Nevertheless, we plan to alleviate this dataset imbalance in our future work by curating a larger, high-quality labeled dataset by reproducing our coding methodology, and/or with the generation of synthetic, balanced datasets. Though we predict that a larger dataset will still have varying proportions of certain influence cues, it will facilitate machine learning with a larger volume of data points.

Additionally, our dataset is U.S.-centric, identified as a limitation in some prior work (e.g., Fletcher et al., 2018; Newman et al., 2019; Kalsnes and Larsson, 2021). All texts were ensured to be in the English language and all three groups of data were presumably aimed at an American audience. Therefore, we plan future work to test Lumen in different cultural contexts.

5.2.2. ML framework

Lumen, as a learning framework, has three main limitations. First, although the two-level architecture provides high degree of flexibility and is general enough to include other predictive features in the future, it also introduces complexity and overhead because tuning the hyper-parameters and training the model will be more computationally expensive.

Second, topic modeling, a key component of Lumen, generally requires a large number of documents of a certain length (usually thousands of documents and hundreds of words in each document, such as a collection of scientific paper abstracts) for topic inference. This will limit Lumen's effectiveness on short texts or when the training data is limited.

Third, some overlap between the LIWC influence features and emotional salience might exist (e.g., the *sad* LIWC category may correlate with the negative emotional salience), which may negatively impact the prediction performance of the machine learning algorithm used in Lumen. In other words, correlation of input features makes machine learning algorithms hard to converge in general.

6. Conclusion

In this paper, we introduced Lumen, a learning-based framework to expose influence cues in text by combining topic modeling, LIWC, sentiment analysis, and machine learning in a two-layer hierarchical architecture. Lumen was trained and tested with a newly developed dataset of 2,771 total texts manually labeled according to the influence cues applied to the text. Quantitative analysis of the dataset showed that *authority* was the most prevalent influence cue, followed by *subjectivity* and *commitment*; *gain* framing was most prevalent in phishing emails, and use of emphasis commonly occurred in fake, partisan, and mainstream news articles. Lumen presented comparable performance with LSTM in terms of *F*1-micro score, but better interpretability, providing insights of feature importance. Our results highlight the promise of ML to expose influence cues in text with the goal of application in tools to improve the accuracy of human detection of cyber-social engineering threats, potentially triggering users to think analytically. We advocate that the next generation of interventions to mitigate deception expose influence to users, complementing automatic detection to address new deceptive campaigns and improve user decision-making when confronted with potentially suspicious text.

Data availability statement

The datasets presented in this study can be found in https://github.com/danielaoliveira/Potentiam.

Author contributions

HS: substantial contributions to the conception or design of the work, analysis of data for the work, drafting, and revising. MS: substantial contributions to the conception or design of the work, acquisition, analysis, and interpretation of data, drafting, and revising. LG: acquisition, analysis, and interpretation of data, drafting, and revising. DC and LC: acquisition of data. JF: acquisition of data, drafting, and revising. DO: substantial contributions to the conception or design of the work, drafting, and revising. All authors contributed to the article and approved the submitted version.

Funding

This work was support by the University of Florida Seed Fund award P0175721 and by the National Science Foundation under Grant No. 2028734. This material was based upon work supported by (while serving at) the National Science Foundation.

Acknowledgments

The authors would like to thank the coders for having helped with the labeling of the influences cues in our dataset.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomp.2022.929515/full#supplementary-material

References

(2019b). *Alerts and Notifications, Information Technology*. University of Pittsburgh.

(2019c). *Phishing Scams Targeting the UMN*. University of Minnesota.

(2019d). *Phish Bowl/Phishing Scams*. UCLA IT Services.

(2019e). Office of Information Security. Phishing. Pennsylvania State University.

(2019f). *Recent Phishing Examples, Library and Technology Services*. Lehigh University.

(2019g). *Phishing Alerts, UA Security*. University of Arizona.

(2019h). *Phishes and Scams*. University of Michigan.

(2021). *Fact Check: Courts Have Dismissed Multiple Lawsuits of Alleged Electoral Fraud Presented by Trump Campaign*. Reuters Staff.

Akbar, N. (2014). *Analysing Persuasion Principles in Phishing Emails*. University of Twente.

Akhawe, D., and Felt, A. P. (2013). "Alice in warningland: a large-scale field study of browser security warning effectiveness," in *22nd USENIX Security Symposium*, 257–272.

Ariely, D., Gneezy, U., Loewenstein, G., and Mazar, N. (2009). Large stakes and big mistakes. *Rev. Econ. Stud.* 76, 451–469. doi: 10.1111/j.1467-937X.2009.00534.x

Bago, B., Rand, D. G., and Pennycook, G. (2020). Fake news, fast and slow: deliberation reduces belief in false (but not true) news headlines. *J. Exp. Psychol. Gen.* 149, 1608–1613. doi: 10.1037/xge0000729

Barnidge, M., and Peacock, C. (2019). A third wave of selective exposure research? The challenges posed by hyperpartisan news on social media. *Media Commun.* 7, 4–7. doi: 10.17645/mac.v7i3.2257

Basnet, R., Mukkamala, S., and Sung, A. H. (2008). "Detection of phishing attacks: a machine learning approach," in *Soft Computing Applications in Industry*, ed B. Prasad (Berlin; Heidelberg: Springer), 373–383. doi: 10.1007/978-3-540-77465-5_19

Blei, D. M., Jordan, M. I., and Ng, A. Y. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.* 3, 993–1022. doi: 10.5555/944919.944937

Blei, D. M., and Lafferty, J. D. (2009). "Topic models, in *Text Mining: Classification, Clustering, and* Applications, eds A. N. Srivastava and M. Sahami (Cambridge, MA: Chapman and Hall/CRC), 71–89. doi: 10.1201/9781420059458.ch4

Bursztein, E., and Oliveira, D. (2019). "Deconstructing the phishing campaigns that target gmail users," in *BlackHat 2019*.

Chandrasekaran, M., Narayanan, K., and Upadhyaya, S. (2006). "Phishing email detection based on structural properties," in *NYS Cyber Security Conference* (Albany, NY).

Cialdini, R. B. (2001). The science of persuasion. *Sci. Am.* 284, 76–81. doi: 10.1038/scientificamerican0201-76

Cialdini, R. B. (1993). *Influence: The Psychology of Persuasion*. New York, NY: Morrow.

Dong, Z., Kapadia, A., Blythe, J., and Camp, L. J. (2015). "Beyond the lock icon: real-time detection of phishing websites using public key certificates," in *2015 APWG Symposium on Electronic Crime Research*, 1–12. doi: 10.1109/ECRIME.2015.7120795

Felt, A. P., Ainslie, A., Reeder, R. W., Consolvo, S., Thyagaraja, S., Bettes, A., et al. (2015). "Improving SSL warnings: comprehension and adherence," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. CHI '15* (New York, NY: ACM), 2893–2902. doi: 10.1145/2702123.2702442

Ferreira, A., and Teles, S. (2019). Persuasion: how phishing emails can influence users and bypass security measures. *Int. J. Hum. Comput. Stud.* 125, 19–31. doi: 10.1016/j.ijhcs.2018.12.004

Fletcher, R., Cornia, A., Graves, L., and Nielsen, R. K. (2018). *Measuring the Reach of "Fake News" and Online Disinformation in Europe*. University of Oxford: Reuters Institute for the Study of Journalism.

Fuller, R. P., and Rice, R. E. (2014). Lights, camera, conflict: newspaper framing of the 2008 screen actors guild negotiations. *J. Mass Commun. Q.* 91, 326–343. doi: 10.1177/1077699014527455

Graham, J., Haidt, J., and Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *J. Pers. Soc. Psychol.* 96, 1029–1046. doi: 10.1037/a0015141

Griffiths, T. L., and Steyvers, M. (2004). Finding scientific topics. *Proc. Natl. Acad. Sci. U.S.A.* 101(Suppl), 5228–5235. doi: 10.1073/pnas.0307752101

Hove, T., Paek, H. J., Isaacson, T., and Cole, R. T. (2013). Newspaper portrayals of child abuse: frequency of coverage and frames of the issue. *Mass Commun. Soc.* 16, 89–108. doi: 10.1080/15205436.2011.632105

Hutto, C. J., and Gilbert, E. (2014). "Vader: a parsimonious rule-based model for sentiment analysis of social media text," in *Eighth International AAAI Conference on Weblogs and Social Media*.

Kahneman, D., and Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica* 47, 263. doi: 10.2307/1914185

Kahneman, D., and Tversky, A. (1980). Prospect theory: an analysis of decision under risk. *Econometrica* 12, 263–292/

Kalsnes, B., and Larsson, A. O. (2021). Facebook news use during the 2017 Norwegian elections–assessing the influence of hyperpartisan news. *J. Pract.*15, 209–225. doi: 10.1080/17512786.2019.1704426

Kircanski, K., Notthoff, N., DeLiema, M., Samanez-Larkin, G. R., Shadel, D., Mottola, G., et al. (2018). Emotional arousal may increase susceptibility to fraud in older and younger adults. *Psychol, Aging*. 33, 325–337. doi: 10.1037/pag0000228

Kühberger, A. (1998). The influence of framing on risky decisions: a meta-analysis. *Organ. Behav. Hum. Decis. Process.* 75, 23–55. doi: 10.1006/obhd.1998.2781

Lawson, P., Zielinska, O., Pearson, C., and Mayhorn, C. B. (2017). Interaction of personality and persuasion tactics in email phishing attacks. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* 61, 1331–1333. doi: 10.1177/1541931213601815

Lim, C. (2018). Checking how fact-checkers check. *Res. Polit.* 5. doi: 10.1177/2053168018786848

Lin, T., Capecci, D. E., Ellis, D. M., Rocha, H. A., Dommaraju, S., Oliveira, D. S., et al. (2019). Susceptibility to spear-phishing emails: effects of internet user demographics and email content. *ACM Trans. Comput. Hum. Interact.* 26, 32:1–32:28. doi: 10.1145/3336141

Lombard, M., Snyder-Duch, J., and Bracken, C. C. (2002). Content analysis in mass communication: assessment and reporting of intercoder reliability. *Hum. Commun. Res.* 28, 587–604. doi: 10.1111/j.1468-2958.2002.tb00826.x

Morey, A. C., and Eveland, W. P. Jr. (2016). Measures of political talk frequency: assessing reliability and meaning. *Commun. Methods Meas.* 10, 51–68. doi: 10.1080/19312458.2015.1118448

Mueller, R. S. (2019). *Report on the Investigation Into Russian Interference in the 2016 Presidential Election*. U.S. Department of Justice Washington, DC.

Naili, M., Chaibi, A. H., and Ghezala, H. H. B. (2017). Comparative study of word embedding methods in topic segmentation. *Proc. Comput. Sci.* 112, 340–349. doi: 10.1016/j.procs.2017.08.009

Newman, N., Fletcher, R., Kalogeropoulos, A., and Nielsen, R. K. (2019). *Reuters Institute Digital News Report 2019*. Reuters Institute.

NLTK (2020). *Natural Language Toolkit*. Available online at: https://www.nltk.org/

Oest, A., Safaei, Y., Doupé, A., Ahn, G. J., Wardman, B., and Tyers, K. (2019). "PhishFarm: a scalable framework for measuring the effectiveness of evasion techniques against browser phishing blacklists," in *2019 IEEE Symposium on Security and Privacy*. doi: 10.1109/SP.2019.00049

Oliveira, D., Rocha, H., Yang, H., Ellis, D., Dommaraju, S., Muradoglu, M., et al. (2017). "Dissecting spear phishing emails for older vs young adults: on the interplay of weapons of influence and life domains in predicting susceptibility to phishing," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. CHI '17* (New York, NY: ACM), 6412–6424. doi: 10.1145/3025453.3025831

Oliveira, D. S., Lin, T., Rocha, H., Ellis, D., Dommaraju, S., Yang, H., et al. (2019). Empirical analysis of weapons of influence, life domains, and demographic-targeting in modern spam–an age-comparative perspective. *Crime Sci.* 8, 3. doi: 10.1186/s40163-019-0098-8

Peace, K. A., and Sinclair, S. M. (2012). Cold-blooded lie catchers? An investigation of psychopathy, emotional processing, and deception detection: psychopathy and deception detection. *Legal Criminol. Psychol.* 17, 177–191. doi: 10.1348/135532510X524789

Peng, T., Harris, I., and Sawa, Y. (2018). "Detecting phishing attacks using natural language processing and machine learning," in *2018 IEEE 12th International Conference on Semantic Computing*, 300–301. doi: 10.1109/ICSC.2018.00056

Pennebaker, J. W., Boyd, R. L., Jordan, K., and Blackburn, K. (2015). *The Development and Psychometric Properties of LIWC*. Univerity of Texas.

Pennycook, G., and Rand, D. G. (2021). The psychology of fake news. *Trends Cogn. Sci*. 25, 388–402. doi: 10.1016/j.tics.2021.02.007

Perreault, W. D. Jr., and Leigh, L. E. (1989). Reliability of nominal data based on qualitative judgments. *J. Market. Res*. 26, 135–148. doi: 10.1177/002224378902600201

Ramage, D., Hall, D., Nallapati, R., and Manning, C. D. (2009). "Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 248–256. doi: 10.3115/1699510.1699543

Rice, R. E., Gustafson, A., and Hoffman, Z. (2018). Frequent but accurate: a closer look at uncertainty and opinion divergence in climate change print news. *Environ. Commun*. 12, 301–321. doi: 10.1080/17524032.2018.1430046

Ross, R. M., Rand, D. G., and Pennycook, G. (2021). Beyond "fake news": analytic thinking and the detection of false and hyperpartisan news headlines. *Judgement Decis. Mak*. 16, 484–504. doi: 10.31234/osf.io/cgsx6

Rothman, A. J., and Salovey, P. (1997). Shaping perceptions to motivate healthy behavior: the role of message framing. *Psychol. Bull*.121, 3–19. doi: 10.1037/0033-2909.121.1.3

Russell, J. A. (1980). A circumplex model of affect. *J. Pers. Soc. Psychol*. 39, 1161–1178. doi: 10.1037/h0077714

Sadeghi, F., Bidgoly, A. J., and Amirkhani, H. (2020). *FNID: Fake News Inference Dataset*. IEEE Dataport.

Sheng, S., Magnien, B., Kumaraguru, P., Acquisti, A., Cranor, L. F., Hong, J., et al. (2007). "Anti-phishing phil: the design and evaluation of a game that teaches people not to fall for phish," in *Proceedings of the 3rd Symposium on Usable Privacy and Security* (New York, NY: ACM), 88–99. doi: 10.1145/1280680.1280692

Shi, H., Gerlach, M., Diersen, I., Downey, D., and Amaral, L. (2019). "A new evaluation framework for topic modeling algorithms based on synthetic corpora," in *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Shyni, C. E., Sarju, S., and Swamynathan, S. (2016). A multi-classifier based prediction model for phishing emails detection using topic modeling, named entity recognition and image processing. *Circ. Syst*. 7, 2507–2520. doi: 10.4236/cs.2016.79217

Smiles, M. (2019a). *Phishing Scam Reports Archive*. Miller Smiles.

Smith, B. L. (2021). "Propaganda," in *Encyclopedia Britannica*. Available online at: https://www.britannica.com/topic/propaganda

Stajano, F., and Wilson, P. (2011). Understanding scam victims: seven principles for systems security. *Commun. ACM* 54, 70–75. doi: 10.1145/1897852.1897872

Steyvers, M., and Griffiths, T. (2007). Probabilistic topic models. *Handb. Latent Seman. Anal*. 427, 424–440.

Stroud, N. J. (2014). *Selective Exposure Theories. The Oxford Handbook of Political Communication*. doi: 10.1093/oxfordhb/9780199793471.013.009

Sunshine, J., Egelman, S., Almuhimedi, H., Atri, N., and Cranor, L. F. (2009). "Crying wolf: an empirical study of SSL warning effectiveness," in *USENIX Security Symposium*, 399–416.

Tausczik, Y. R., and Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *J. Lang. Soc. Psychol*. 29, 24–54. doi: 10.1177/0261927X09351676

Twitter Help Center. *Government and State-Affiliated Media Account Labels*. Available online at: https://help.twitter.com/en/rules-and-policies/state-affiliated

U.S. House of Representatives Permanent Selection Committee on Intelligence (0000). *Social Media Advertisements*. Available online at: https://intelligence.house.gov/social-media-content/social-media-advertisements.htm

van der Heijden, A., and Allodi, L. (2019). "Cognitive triaging of phishing attacks," in *28th USENIX Security Symposium* (Santa Clara, CA: USENIX Association), 1309–1326.

Vance, A., Kirwan, B., Bjornn, D., Jenkins, J., and Anderson, B. B. (2017). "What do we really know about how habituation to warnings occurs over time?: A longitudinal fMRI study of habituation and polymorphic warnings," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (New York, NY: ACM), 2215–2227. doi: 10.1145/3025453.3025896

Wright, R. T., Jensen, M. L., Thatcher, J. B., Dinger, M., and Marett, K. (2014). Influence techniques in phishing attacks: an examination of vulnerability and resistance. *Inform. Syst. Res*. 25, 385–400. doi: 10.1287/isre.2014.0522

Xu, W. W., Sang, Y., and Kim, C. (2020). What drives hyper-partisan news sharing: exploring the role of source, style, and content. *Digit. J*. 8, 486–505. doi: 10.1080/21670811.2020.1761264

Zhou, X., and Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Comput. Surv*. 53, 1–40. doi: 10.1145/3395046