Posts of Peril: Detecting Information About Hazards in Text

Keith Burghardt[1], Daniel M.T. Fessler[2,3,4], Chyna Tang [2,5], Anne Pisor [6], Kristina Lerman[7],
[1] School of Data Science, University of North Carolina at Charlotte, Charlotte, NC 28223, USA
[2]Department of Anthropology, University of California, Los Angeles, Los Angeles, CA 90095, USA
[3]Bedari Kindness Institute, University of California, Los Angeles, Los Angeles, CA 90095, USA
[4]Center for Behavior, Evolution, & Culture, University of California, Los Angeles, Los Angeles, CA 90095, USA
[5]Department of Psychology, University of California, San Diego, San Diego, CA 92093, USA
[6]Department of Anthropology, The Pennsylvania State University, University Park, PA 16802, USA
[7]Luddy School of Informatics, Indiana University, Bloomington, IN 47408, USA
kburghar@charlotte.edu, dfessler@anthro.ucla.edu, pisor@psu.edu, chynst@g.ucla.edu, krlerman@iu.edu

Abstract

Socio-linguistic indicators of affectively-relevant phenomena, such as emotion or sentiment, are often extracted from text to better understand features of human-computer interactions, including on social media. However, an indicator that is often overlooked is the presence or absence of information concerning harms or hazards. Detecting such indicators in text is important because substantial research demonstrates that negative events are more likely to be attended to, and more likely to elicit a response. In addition, statements about hazards are often found to be more believable than statements about benefits. Here, we develop a new model to detect information concerning hazards, trained on a new collection of annotated X posts. We show that not only does this model perform well (outperforming, e.g., dictionary approaches), but that the hazard information it extracts is not strongly correlated with such widely used indicators as moral outrage, sentiment, and emotions. (That said, in accord with expectations, hazard information does correlate positively with such emotions as fear, and negatively with emotions like joy.) To demonstrate the utility of our tool, we apply it to two datasets of X posts that discuss important geopolitical events, namely the Israel-Hamas war and the 2022 French national election. In both cases, we find that hazard information, especially information concerning conflict, is common. We extract accounts associated with information campaigns from each data set to explore how information about hazards could be used to attempt to influence geopolitical events. We find that inorganic accounts representing the viewpoints of weaker sides in a conflict often discuss hazards to civilians, potentially as a way to elicit aid for the weaker side. Moreover, the rate at which these hazards are mentioned differs markedly from organic accounts, likely reflecting information operators' efforts to frame the given geopolitical event for strategic purposes. These results are first steps towards exploring hazards within an information warfare environment. The model is shared as a Python package to help researchers and journalists analyze hazard content.

Code — https://github.com/KeithBurghardt/HazardPackage

Datasets — https://github.com/KeithBurghardt/HazardPackage/tree/main/ground_truth_data

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Introduction

Because the mind treats information regarding hazards differently than other types of information, developing techniques to identify hazard information can shed important light on the impact of online content. Humans pay particular attention to information that could signal potential hazards or dangers in their environment (Soroka, Fournier, and Nir 2019; Öhman and Mineka 2001). As a result, negative events tend to have more potent and lasting effects than their positive counterparts (Baumeister et al. 2001; Rozin and Royzman 2001). In a similar vein, negative news attracts more attention (Soroka, Fournier, and Nir 2019), negative information spreads more readily than positive information (Bebbington et al. 2017; Blaine and Boyer 2018; Fay et al. 2021; Ferrara and Yang 2015; Brady et al. 2017), and fear speech (negative information about a particular group) can drive groups towards conflict (Saha et al. 2023). Compounding overall negativity bias, people tend to find information about hazards more believable than information about benefits, a property termed negatively-biased credulity (Fessler, Pisor, and Navarrete 2014). In accord with these features of negative information, such information is a factor in the spread of misinformation (Vosoughi, Roy, and Aral 2018; Ecker et al. 2022; Youngblood et al. 2023), suggesting that it could be a potent tool for information operations, i.e., sets of accounts aimed at influencing public opinion online.

Although prior research has developed vital tools for identifying a variety of dimensions of social media content, to date, the importance of extracting information concerning hazards has been largely overlooked. We define hazards as negative events that entail a high probability of a significant (and often immediate) cost to individual interests (Fessler, Pisor, and Navarrete 2014). Although not all hazards are life-threatening or externally imposed (e.g., natural disasters such as hurricanes), they all represent threats that warrant attention and response. It is particularly critical to analyze information concerning hazards during events such as terror attacks or wars (Choi et al. 2022; Jamil, Pais, and Cordeiro 2022; Alshehri, Abdul-Mageed et al. 2020), where the dissemination of threat-related language can significantly influence public perception and behavior.

We address the gap in the existing literature by devel-

oping a multilingual model to recognize information concerning hazards (hereafter, for simplicity, "hazards") expressed in text messages on X (formerly Twitter). We validate the model on human-annotated data and show substantially higher performance over baselines. The model is highly scalable, enabling us to quantify the likelihood that hazards are present in millions of posts. We show that while the model's hazard classifier is somewhat correlated with negative emotions such as anger and fear, it goes beyond such indices, capturing additional salient information about potential dangers.

We apply this model to two diverse multilingual X datasets that serve as case studies of the model's utility. Namely, we detect hazards in posts from 2023 about the Israel-Hamas war, and posts from 2022 about the French national election. The datasets are unique, from different locations and types of conflict (war versus political contest), as well as different languages, including Arabic, French, and Chinese. Moreover, these data contain material that we attribute to information operations. We see notable differences in how information operation accounts describe threats (for example, emphasizing harms in Gaza and not in Israel), while also using hazards to demean an opponent. We also find that the prevalence or frequency of specific hazards differs greatly between different information operations, reflecting different interest factions within each dataset. These findings appear to show that hazards to civilians are consistently used to justify aid to the weaker side in a conflict. More broadly, these results demonstrate how identifying hazards can reveal salient features, potentially unmasking information operation tactics. Because our hazard-detection model has broad utility, we have created a package that enables any researcher to apply it to their own datasets; we do so recognizing that the present effort is a first step towards comprehensive hazard detection, as additional research will improve the model.

In summary, our contributions are as follows.

- We develop and share a new model to extract hazards from social media posts.
- To demonstrate the model's utility, we apply it to two newly collected online datasets, and contrast hazards to alternative indicators of affect and threat.
- We analyze hazards used in information operations to further reveal influence tactics.
- We share hazard-annotated data from a wide range of X posts so as to enable researchers to iteratively improve upon this model.

Overall, these results demonstrate the need for, and utility of, our new tool to detect information about hazards, especially in social media environments.

Related Work

Information Operations

Our work focuses on hazards in information operations, coordinated efforts to change behaviors or opinions (Burghardt, Hogg, and Lerman 2018), especially during important events, such as the Israel-Hamas war (Dey, Luceri, and Ferrara 2024), or elections (Badawy, Ferrara, and Lerman 2018; Burghardt, Hogg, and Lerman 2018). Detecting information operations is difficult, as, with few exceptions (e.g., (Luceri et al. 2024)), investigators generally lack access to ground-truth datasets with which to train or evaluate a model. Instead, researchers typically aim to detect strong coordination, such as unusually similar text or hashtags (Burghardt, Hogg, and Lerman 2018; Luceri et al. 2024), or quick reposts (Mazza et al. 2019). In the present study, we use hashtag matching (Burghardt, Hogg, and Lerman 2018; Luceri et al. 2024), as well as a method by Luceri et al. (Luceri et al. 2024) which combines several different approaches together: co-posting, co-URL sharing, fast reposting, hashtag matching, and text similarity. It then finds the eigenvector centrality of the network, as high eigenvector central nodes are more likely to be part of an information campaign.

Linguistic Indicators in Text

Detecting indicators from text has a long history. One of the first methods to do this was The General Inquirer (Stone, Dunphy, and Smith 1966), and, more recently, text indicators have been analyzed with LIWC (Pennebaker, Francis, and Booth 2001). These dictionary-based approaches have subsequently been largely replaced by rule-and-word approaches, such as VADER (Hutto and Gilbert 2014), which have themselves been superseded by text embedding approaches, including SeerNet (Duppada, Jain, and Hiray 2018), a moral outrage classifier (Brady et al. 2021), and DeepMoji (Felbo et al. 2017). The advantages of these approaches are that embeddings can learn how semantically similar content has similar indicator values (e.g., "I am sad" and "This day was awful" would be viewed as sad statements even when the two sentences do not have any words in common). These approaches are based on GRU (Dey and Salem 2017) or LSTM (Hochreiter and Schmidhuber 1997); however, more recently, transformer-based models (Vaswani et al. 2017), such as BERT (Devlin et al. 2018) and Sentence-BERT (Reimers and Gurevych 2019), have become popular for uses such as detecting hate speech (Davani, Díaz, and Prabhakaran 2022) or emotions (Acheampong, Nunoo-Mensah, and Chen 2021). Unlike prior approaches, transformer models account for context. Finally, these methods have been improved with SpanEmo (Alhuzali and Ananiadou 2021) or Demux (Chochlakis et al. 2023), which account for correlations between labels.

LLM Approaches to Text Analysis

While traditional methods to detect linguistic indicators are either dictionary approaches, e.g., LIWC (Pennebaker, Francis, and Booth 2001), or supervised methods on smaller language models, such as Demux (Chochlakis et al. 2023), there has been a recent advance in LLM appropaches, especially given their wide applicability (Cao et al. 2025). These approaches include using Large Language Models (LLMs) to similarly detect emotion (Peng et al. 2024) or othering language (Gerard, Weninger, and Lerman 2025) via LoRA fine-tuning (Hu et al. 2022) and Prompt Tuning (Liu et al. 2021). These techniques (along with QLoRA

(Dettmers et al. 2023)) allow for efficient fine-tuning of multi-billion parameter models by adding a low-rank matrix to each layer. Alternative methods include few-shot learning (Hong et al. 2025), in which prompts are injected with additional examples to give context to a prompt. In contrast, reasoning is found to improve model performance, and therefore having LLMs explain their step-by-step reasoning improves their performance across a range of tasks (Hama, Otsuka, and Ishii 2024).

Here, we develop a transformer-based model to extract hazards–a previously under-studied indicator–from social media posts. We apply a range of models to sentence embeddings and compare these with both LLM (GPT-3.5, GPT-4, and GPT-5 (Achiam et al. 2023)) and dictionary approaches to detect hazards in text (Choi et al. 2022). As effective baselines, we include both few-shot prompting and reasoning, but leave a fine-tuned LLM baseline as future work. While Choi et al.'s analysis of threats in text is perhaps most similar to our project, we detect hazards rather than threats. Moreover, we apply multi-lingual AI models to detect hazards, which we find is a significant improvement over dictionary approaches, such as that of Choi et al. (Choi et al. 2022). This work also compliments prior work detecting hazardous events in a time series (Jamil, Pais, and Cordeiro 2022), as well as explicit threats (dangerous speech) (Alshehri, Abdul-Mageed et al. 2020).

Negatively-Biased Credulity

As noted above, a growing corpus of research documents that negative information spreads more readily than positive information. For example, negative content is more likely to be shared (Martel, Pennycook, and Rand 2020; Youngblood et al. 2023); moral-emotional language (often in large part negative in nature) spreads the most in partisan discussions (Brady et al. 2017); negative sentiment posts spread faster (Ferrara and Yang 2015); and moral outrage makes posts more viral (Brady et al. 2021). One factor likely contributing to the asymmetry in the spread of negative versus positive information is that negative information – particularly information about potential dangers – is more likely to be believed. For people past and present, believing false information about hazards has, likely been less costly on average than rejecting true information about hazards, since taking unnecessary precautions generally entails less dire consequences than does failing to take necessary precautions against threats; in contrast, there likely has not been an overarching pattern with regard to the respective costs and benefits of believing or not believing information about benefits. People therefore broadly find information about hazards more believable than information concerning benefits, a pattern termed negatively-biased credulity (Fessler, Pisor, and Navarrete 2014; Fessler, Pisor, and Holbrook 2017; Fessler 2019; Samore et al. 2018; Forgas 2019). Of particular relevance in regard to information operations, work on negatively-biased credulity articulates with investigations on the role of credulity in recipients' susceptibility to manipulation (Little 2017; Kartik, Ottaviani, and Squintani 2007).

Although there are many points of contact between psychological research on negatively-biased credulity and existing work on the spread of information, misinformation, and disinformation, this construct also calls attention to potentially important distinctions that have been overlooked in prior research on information spread. With a logic grounded in evolutionary theory, a core insight of work on negatively-biased credulity is that not all negative information confers a similar survival advantage. Specifically, swiftly recognizing imminent threats (Öhman and Mineka 2001) (marked, for example, by fear) is more critical to survival than is revisiting past losses (marked, for example, by sadness). Hence, while negative emotions as a category may reduce belief in a false claim (Phillips et al. 2025), discussion of hazards seems to increase it (Fessler 2019). Likewise, although other evidence shows that overall emotion perception is associated with both poor misinformation discernment and misinformation sharing (Bago et al. 2021; Martel, Pennycook, and Rand 2020), such work does not differentiate between emotions attending messages concerning hazards and those attending messages concerning benefits. As a metric of content, it is therefore critical to distinguish information about hazards from negative valence in general, the presence of emotion-laden information, or other features of language.

We expand on previous research regarding negativity by developing a novel hazard detection tool. We demonstrate this tool's utility by applying it to millions of social media posts. We employ large datasets that provide enough statistical power for us to assess how statements about hazards vary after major events, how they relate to other indicators, and how diverse groups discuss hazards. In turn, this can inform hypotheses about how hazard information is harnessed when trying to influence social media users.

Research Methods

All data collected and analyzed were determined to be exempt from assessment by the institutional review board of the lead author's university, where all modeling, data collection and data analysis were conducted. All annotations were non-human-subject research. In addition, all data were anonymized prior to analysis or annotation to minimize privacy risks. This was accomplished via anonymization of user names and profile content.

Hazards Benchmark: Data and Model

We curate a ground truth dataset with which to train models to recognize hazards. This process involves collecting and annotating posts for the presence of hazards. This ground truth dataset is then used to train a language model to classify hazards posts.

Ground Truth Data To create the benchmark X post dataset, we first extracted 1,338 X posts containing at least one word from the Threat Dictionary (Choi et al. 2022). Although we are not aware of any dataset annotated for hazards, we chose this size as approximately the median number of posts annotated for valence across several previous datasets (cf. Table 1 in (Mendes and Martins 2023)). We confirm that this is a sufficient sample through the model performance described in the Results section. In order to produce a representative sample, data are collected via X's

Academic API between March 2006 (when X, then Twitter, was founded) and late 2022.

We randomize the order of these posts and recruit Cloud Research annotators to label any hazards present therein via a Qualtrics survey. For each post, workers answer, "Does the tweet describe a hazard (something that could impose harm or other costs on the author of the tweet or on others)?". (Our annotations predate X's name change from Twitter, hence our use of "tweet" in the annotation question rather than "post".) Workers are paid $2 for each assignment in which they annotate 10 random posts (on average this took 11 minutes to complete, hence compensation was equivalent to $12 an hour). To account for workers who do not meaningfully complete the assignment, we add an easy question (specifically "choose the answer to 2+2") midway through the survey and remove any workers who did not complete it. We also remove annotations that are unfinished or take less than 200 seconds to complete (this was an arbitrary cutoff to better ensure that the annotations were not completed carelessly). As all data are annotated before 2023, we believe that the prevalence of workers annotating using LLMs (Veselovsky, Ribeiro, and West 2023) was minimal.

To check inter-rater reliability, we used the R library `irrNA` (Brueckl and Heuer 2022), which assumes randomly missing data (this assumption is consistent with our annotation methodology, as we assign 10 posts at random to each rater). Applied to the dataset, we have an Intraclass Correlation Coefficient of: ICC(1) = 0.12, ICC(k) = 0.29, agreement of ICC(A,1) = 0.17, ICC(A,k) = 0.37, ICC(C,1)=0.17, and ICC(C,k) =0.37 (p-values < 0.001). These values can be interpreted as moderate consistency but poor agreement. However, these values are in keeping with other subjective text rating tasks, including hate speech (Sachdeva et al. 2022) and emotions (Mohammad et al. 2018). The low ICC may also reflect the difficult nature of the task that we gave annotators, as they were trying to distinguish between posts with versus without hazards despite the fact that all of the annotated posts contain Threat Dictionary words.

All posts annotated by fewer than two crowd workers are discarded, resulting in a dataset with 1,131 posts for training, validation, and testing. We use Python's demoji library (https://pypi.org/project/demoji/) to convert all emojis to words in order to reduce artifacts in the embeddings. All data are collected according to the FAIR principles: Findable (these data are directly available via the repository link at the end of the Introduction section, and contain a unique identifier with all metadata described), Accessible (the link is accessible to anyone), Interoperable (metadata use a formal and broadly applicable language), and Reusable (all data are described in detail). In the repository link, we also include a datasheet for the dataset Gebru et al. (2021).

Additional Dataset Annotations Random posts that we use to train the model might not be representative of the Israel-Hamas war and 2022 French election datasets to which the model is to be applied. To address this potential limitation, we trained three students to annotate 150 random posts within each of the respective datasets. To adequately compare our model against the Threat Dictionary, and because these students were native English speakers, we translated all text to English prior to annotation. We specifically used X translated text for French posts, and translate Arabic text to English using Google Translate as the Twitter translations were not available for that set of text. We provided each student with the same guidelines given to Cloud Research workers, but to ensure data quality, we also had students label posts "-1" if there was not enough information to determine whether the post contained a hazard (e.g., a post containing a URL), and "1/2" if the annotator had low confidence that the post contains a hazard due to an ambiguous scenario (e.g., parading around a body). To clean data, we removed any post that any of the annotators rated as "-1"; for the rare cases in which a post was labeled "1/2", we designated the post as "not hazard". This makes it especially difficult for the model, as it was trained on tweets containing hazards, and would presumably have a higher false-positive rate. After data cleaning, we have 124 annotated posts in the Israel-Hamas war dataset, and 135 posts in the 2022 French election dataset. The Fleiss Kappa scores for these annotations are 0.40 and 0.49 for the Israel-Hamas and French election datasets, respectively, representing moderate agreement. There were 74% and 10% of data labeled hazards in the respective datasets (which intuitively indicates a high rate of hazard discussions within a war context).

Model Training We use 90% of data for training or validation and 10% for testing, ensuring a sufficient amount of training data for each model. We choose from four multi-lingual sentence transformers: `distiluse-base-multilingual-cased-v2`, `paraphrase-multilingual-MiniLM-L12-v2`, `Qwen3-Embedding-0.6B`, and `stsb-xlm-r-multilingual` to embed text in the datasets. Multilingual text embedding models were used because of the different languages in each social media dataset. We then apply several supervised models to these embeddings, including XGBoost (Chen and Guestrin 2016) (via https://xgboost.readthedocs.io/), neural networks (Abadi et al. 2015), random forest (du Boisberranger 2024a; Ho 1995; Pedregosa et al. 2011), and support vector machines (SVMs) (du Boisberranger 2024b; Hearst et al. 1998; Pedregosa et al. 2011); the latter two are trained via `scikit-learn` (Pedregosa et al. 2011). We use BayesSearchCV to optimize hyperparameters for random forest, SVM, and XGBoost (Head et al. 2024). We find the optimal parameters via five-fold cross-validation of the data split after ten iterations and minimize each model's default scoring metric. We also use GPyOpt Bayesian Optimization for a feedforward neural network (authors 2016). The best hyperparameters are shown in the GitHub repository.

We also experimented with augmenting human annotated posts with 5,000 GPT-3.5 annotated posts that contained words from the Threat Dictionary (Choi et al. 2022), and 5,000 posts collected at random (containing popular English keywords, namely any of the top 100 lemmas within a large corpus (https://www.wordfrequency.info/samples.asp) that X does not consider stop words. This augmentation did

not significantly change the performance of the model. All models are trained via NVIDIA Tesla K80 GPU with 12GB of VRAM; we predict text via GeForce RTX 2080 GPU with 8GB VRAM. For comparison, we also use GPT-3.5, GPT-4, and GPT-5 (Achiam et al. 2023) with few-shot learning and chain-of-thought prompting (Wei et al. 2022) via the prompt, "Does the tweet [story] describe a hazard (something that could impose harm or other costs on the author of the tweet or on others)? Please answer 'yes' or 'no' and explain your thought process." and include zero, two, or five examples of posts containing hazards and posts that do not. All texts with "yes" are labeled "hazard," and those without are labeled "not hazard" (rare situations where the LLMs do not know if a hazard exists in text are labeled "not hazard").

Comparison to Alternative Text Indicators

We compare our hazard detection model to alternative text indicators of affect-related phenomena: moral outrage (Brady et al. 2021), sentiment (VADER) (Hutto and Gilbert 2014), emotion detection (Demux) (Chochlakis et al. 2023), and threat words (Choi et al. 2022). These are state-of-the-art methods for detecting each indicator. These code are under a MIT license (VADER and Demux), and Creative Commons Attribution-NonCommercial-ShareAlike 2.0 license (moral outrage), respectively, and are adapted as needed to run on a GeForce RTX 2080 GPU with 8GB VRAM. All model outputs that we analyze are continuous (such as confidence values for emotions), with the exception of the Threat Dictionary (Choi et al. 2022), in which we indicate if a threat word is (1) or is not (0) present in a post. Because the Threat Dictionary is not necessarily a comprehensive dataset, we also found the top 3 nearest synonyms to each word in the Threat Dictionary using a pruned Word2Vec trained on the Google News Dataset (Bird and Loper 2004). We then compare the performance of this enhanced dictionary on annotated datasets to the best-performing model. We also compare how the enhanced threat dictionary results compare on the two datasets studied. The enhanced dictionary is available on our GitHub.

X Data

To demonstrate the hazard detection model's utility, as case studies, we apply it to two separate datasets.

Israel-Hamas War Our analysis uses a corpus of 3.6M posts, from 1.3M accounts, spanning the period from August 31 to November 1, 2023, about the 2023 Hamas attack on Israel and Israel's subsequent invasion of Gaza. The posts were collected by querying X with a set of keywords related to the war: e.g., "Israel," "Hamas," "Gaza," etc. These data are multilingual, with approximately 93% of the posts in English, 6.5% in Arabic, and a small proportion in other languages.

2022 French Election We also analyze 5.9M posts, from 677K accounts on X, related to the 2022 French election that took place April 10–24, 2022. The dataset spans February 14 to June 30, 2022. The posts were collected by querying X with a set of keywords related to the election, e.g., major candidates such as "Macron", "Marine Le Pen", "Mélenchon", etc. These data were also multilingual, with 95% in French and 5% in English.

Coordination

We further partition each dataset using a common coordination metric (Burghardt et al. 2023; Luceri et al. 2024), in which accounts are deemed coordinated if they post near-duplicate sequences of hashtags (which are strongly associated with near-duplicate messages). More specifically, for any pair of accounts, we label them as coordinated if (a) they both have posts containing three or more hashtags, and (b) the sequence of hashtags is exactly the same between both posts. While simple, this method has proven to be a very effective tool for detecting coordinated accounts (Burghardt et al. 2023). This method is effective because even when coordinated accounts try to obfuscate their authenticity by re-writing a post across various posts, the sequence of hashtags is often the same. In addition, we create a network of accounts that are linked based on this indicator of coordination. Connected clusters often represent distinct sets of messages consistent with particular information operations.

Applying this method to the Israel-Hamas war dataset, we extract 4.2K coordinated accounts that posted 69K posts, of which 82.7% were in English, 16.7% in Arabic, 0.5% in Hebrew, and a small proportion in other languages. For the 2022 French Election dataset, we extract 179 accounts that posted 27K posts, of which 91% were in French and the rest were in English (a substantially smaller percentage that were in French compared to the data at large). To check the robustness of these results, we also used an alternative state-of-the-art coordination algorithm (Luceri et al. 2024), in which we calculated the co-repost similarity, co-url similarity, rapid retweet similarity, text similarity, and the hashtags shared, and combined these into one network. Within this network, we calculated accounts as coordinated if their eigenvector centrality were ¿ 0.002, following prior work that shows even this modest centrality metric strongly separates coordinated accounts from authentic users (Luceri et al. 2024). This method finds 409 accounts that post 6.8K posts in the Israel-Hamas dataset and 1.8K accounts that post 9.2K posts in the French election dataset. In the former dataset, we find 99.6% of coordinated account posts are in English, and the rest in Arabic, while in the latter dataset, 91% of posts are in French and the rest in English.

We share the coordinated networks in the repository, but, because X's terms of service do not allow us to share post text, and because we have removed all text IDs and account IDs, we do not share additional data. Because these data are public and anonymized, it was not necessary to obtain consent from accounts to extract these data. We show the frequency of posts over time for coordinated and authentic (non-coordinated) accounts in the Appendix (Figs. S1).

Results

Hazards Model and Validation

Performance of the hazards detection model on benchmark data is shown in Fig. 1. Despite the simplicity of the XGBoost model, its performance is comparable to LLM mod-

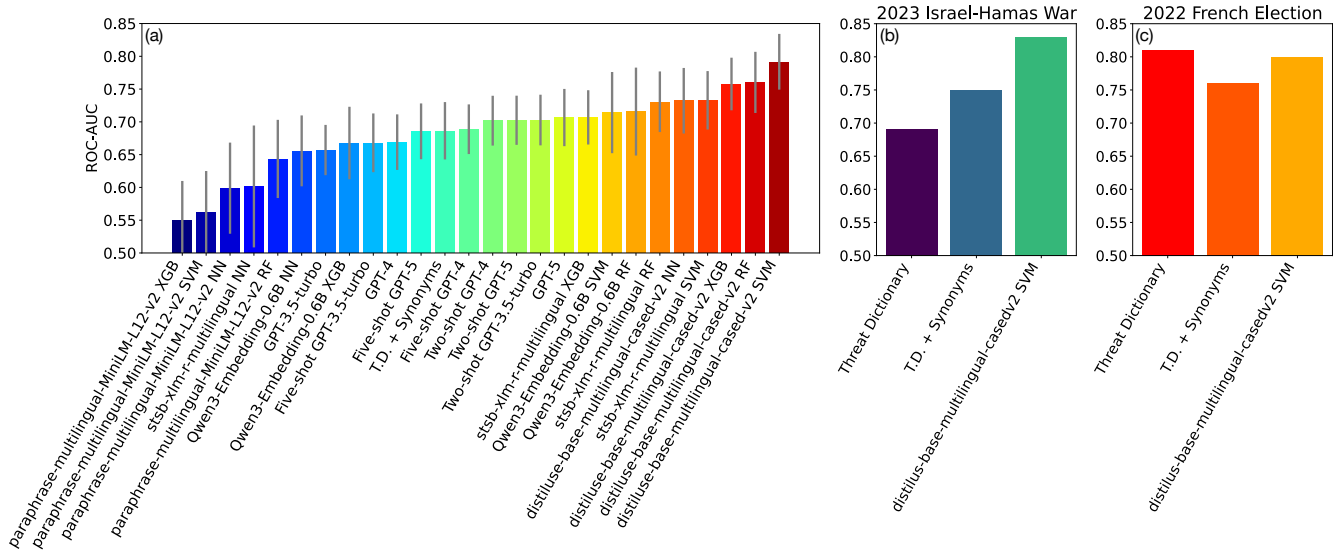

Figure 1: Performance of models on human-annotated X posts. (a) We show the ROC-AUC of XG-Boost (XGB) (Chen and Guestrin 2016), a neutral network (NN) (Abadi et al. 2015), random forest (RF) (Ho 1995), and a support vector machine (SVM) (Hearst et al. 1998) trained on human-annotated posts for multi-lingual embedding models described in the Scientific Methods section. We also show the ROC-AUC of GPT-3.5, GPT-4, and GPT-5 with zero-shot, two-shot, and five-shot predictions. Finally, we also show performance of the Threat Dictionary plus synonyms ("T.D. + Synonyms") at predicting threat text (if a post contained a word that was in this set of words, we labeled the data "hazard"; otherwise it was not). Gray bars represent standard deviations across 50 evaluations. (b) Best model performance and baseline performance on posts from the 2023 Israel-Hamas war dataset. (c) Best model performance and baseline performance on posts from the 2022 French election dataset.

els, with an area under the receiver operating characteristic curve (ROC-AUC) of 0.79 $\pm$0.04 (Fig. 1). Due to the variance in the performance metric (gray bars), it is uncertain whether the model outperforms all models, although its substantially higher throughput and low cost make it an obvious choice compared to GPT models when applied to millions of social media posts. This training dataset uses posts that contain words from the Threat Dictionary (Choi et al. 2022), making the dataset especially challenging, as, despite the posts containing threat words, only a subset include information about hazards. Notably, this implies that simple lexical models, like those based on the Threat Dictionary and synonyms, are a poorer tool, as the Threat Dictionary + Synonym baseline has an ROC-AUC of 0.68 ± 0.04. This demonstrates the need for an AI-based method that can go beyond dictionary baselines. We also compare our human annotations of posts in the Israel-Hamas and 2022 French election dataset, respectively (see details in the Research Methods section), where we translate all text to English to give dictionary-based methods a more even footing. The results indicate the model achieves strong generalizability, even across thematically distinct datasets and different years, and captures context better than dictionary approaches.

Hazards in Real-World Data

We show in Fig. 2 how the hazard confidences compare to other linguistic indicators of affect in our two X datasets on important geopolitical events. We then show the words most often seen in high- and low-hazard posts in each dataset, clarifying how hazards are typically described. Next, we show how hazards are discussed over time, especially within different sets of coordinated accounts that concern each geopolitical event.

Linguistic Analysis of Hazards As we detail in the Discussion section, our model can potentially be used to tackle a variety of questions in which the presence of hazard information is relevant. In order to demonstrate the model's utility as a research tool, below we employ the model to investigate coordinated accounts in social media.

Figure 2 shows correlations between hazard confidences and other text indicators of affect-relevant phenomena expressed in each post, including moral outrage (Brady et al. 2021), sentiment (Hutto and Gilbert 2014), emotion confidence (Chochlakis et al. 2023), and threat words (Choi et al. 2022). In all cases, the absolute value of Spearman correlations is below 0.5, suggesting that alternative indicators do not fully capture information about hazards in text. That said, the positive or negative direction of the correlations makes intuitive sense. For example, moral outrage is weakly positively correlated with hazards, perhaps because people share their moral outrage towards some hazards, such as those that harm the innocent. Similarly, negative sentiment, as well as most negative emotions and posts containing Threat Dictionary (Choi et al. 2022) words, show positive correlations with hazards. This is consistent with the negative framing of the hazard posts (cf. example hazard annotations and predictions in the Appendix Table S1).

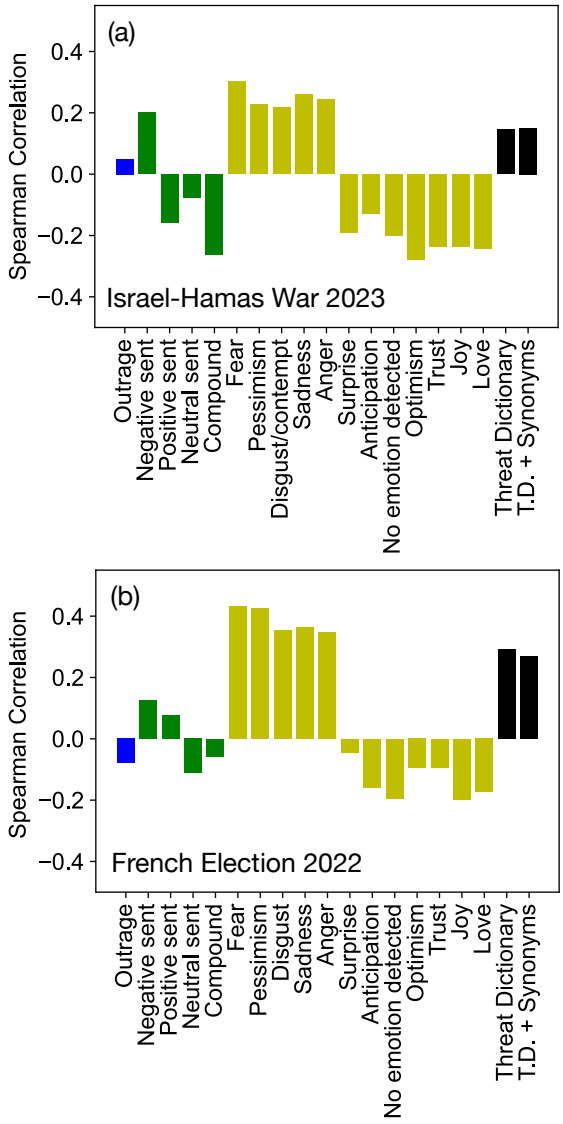

Figure 2: Understanding hazards. Spearman correlation between hazards moral outrage (Brady et al. 2021), sentiment (Hutto and Gilbert 2014), emotions (Chochlakis et al. 2023), Threat Dictionary (Lilienfeld and Latzman 2014) and threat synonyms for (a) the Israel-Hamas war, (b) the 2022 French election. All values are statistically significant (p-value < 0.05).

We also find in Appendix Tables S2 & S3 that high-hazard post words relate to "terror", "bien-être" (well-being), or "corrupt" while low-hazard post words include "pond," "demain" (request), or "2019," which appear in posts that describe non-violent concepts, especially those that are not time-sensitive. While some hazard words are similar to threat words seen in (Choi et al. 2022), some hazard-related words cannot easily be captured with the Threat Dictionary, such as "children," "send," or "American," as these are associated with the target of harms (harm of the innocent, sending aid, etc.).

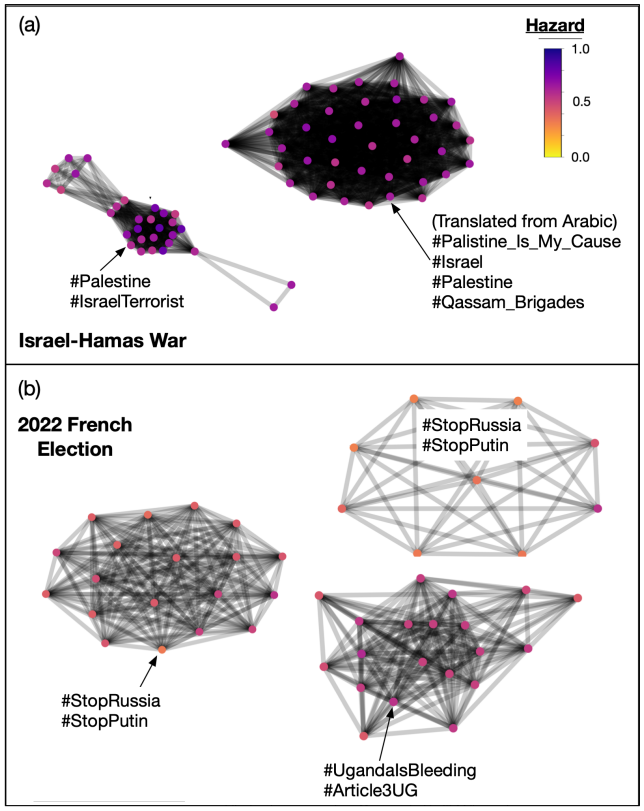

Figure 3: Hazards in representative large coordinated account clusters for (a) the Israel-Hamas war, (b) the 2022 French election. Mean hazard confidence per account is shown as a color from yellow to purple. Top hashtags and example posts are shown next to each cluster.

Israel-Hamas War Next, we explore how hazards are used within coordinated influence campaigns, specifically how hazard content differs between authentic and inauthentic posts in the Israel-Hamas war dataset.

Appendix Fig. S2 shows that likely authentic accounts emphasize the association between hazards and children, civilians, and terror (reflecting hazards surrounding the October 7th Hamas attack on Israel, as well as later bombings of Gaza). In contrast, coordinated accounts do not appear to associate the terrorist attack with a hazard, but instead associate hazards with bombs and civilians (reflecting Israel's assault on Gaza). Analysis of individual posts confirms these observations. We show words associated with high- and low-hazard posts for coordinated and authentic accounts in Appendix Tables S4, S5, S6 & S7. We find that the hazard rate is not statistically significantly different overall between coordinated and authentic accounts (Mann-Whitney U test p-value > 0.1), where both have a mean hazard confidence of 0.54 across all posts (including reposts), but this belies significant differences within clusters and across time. Moreover, with an alternative coordination metric (Luceri

et al. 2024), we find coordinated accounts have significantly higher mean model confidence value (0.55 vs 0.54, Mann-Whitney U test p-value < 0.0001).

We look within coordinated accounts in Fig. 3a, which shows examples of smaller coordinated account clusters. The links connect accounts that are deemed to be coordinated based on their text (see Methods). The top hashtags are shown next to each cluster. Both clusters in this figure show higher hazard confidence (0.55 and 0.63 for the left and right clusters, respectively, versus 0.54 for all non-coordinated clusters, Mann-Whitney U test p-value < 0.001), apparently because of the discussion of hazards against civilians (e.g., "#IsraelTerrorist" is the second-most popular hashtag in the cluster on the left). We show embeddings of posts from these clusters are shown in the Appendix Fig. S4a, where we see the posts have overlapping embedding distributions, suggesting the topic themes are similar (Grootendorst 2022), in agreement with the top hashtags observed.

The top hashtags for the largest cluster of our dataset (1.9K accounts, 42K posts, mean hazard confidence across all posts: 0.52) are '#IsraeliNewNazism', (3.3K posts), '#Gaza_under_attack' (2.3K posts), '#Israel', (2.3K posts), and so on. Posts on October 7, 2023, the day of the Hamas attack, discount the Israeli lives lost with posts such as "Palestinian rockets launched from Gaza hit the center of Israel's capital city, Tel Aviv. Finally Israel Is Going To Be Finished Today Insha'Allah". Posts long after the attack, meanwhile, discuss hazards toward Gazans, e.g., ""My three children. I lost them all!"..." (And interestingly, the same accounts promote protests in the U.S., with posts such as "Enthusiastic demonstration of students and professors of the University of California, Berkeley in support of Palestine..."). In general, these results paint a picture of a selective discussion of hazards.

2022 French Election As a second demonstration of the model's utility, we analyze social media related to the 2022 French election. Coordinated influence campaign accounts appear to hijack election discussions to request aid to Ukraine during the Russia-Ukraine war, as shown in Fig. 3b, where we plot the largest clusters consisting of 48 out of 179 coordinated accounts. Account-level hazard rates do not differ substantially between coordinated and authentic accounts (0.35, Mann-Whitney U test p-value > 0.1), although for an alternative coordination metric (Luceri et al. 2024), we find coordinated accounts have significantly higher confidence (0.48 vs 0.35, Mann-Whitney U test p-value < 0.0001). These findings match what we also see in the Israel-Hamas war dataset. Appendix Fig. S3 shows that likely authentic accounts emphasize the association between hazards and the Russia-Ukraine war ("russie" and "l'ukraine"), while coordinated accounts discuss nuclear ("nucléaire") or associate Russia with hazards "russian" vs low-hazard "l'ukraine". See also the words most, and least, associated with hazard posts in Appendix Tables S8, S9, S10 & S11. To better understand these results, we analyze the individual clusters of coordinated accounts.

In Fig. 3b, we see 3 major clusters. The second and third-largest of which (28 accounts, 361 posts, mean hazard confidence across all posts: 0.53 for the largest, and 0.45 for the smallest cluster, Mann-Whitney p-value < 0.001, compared to non-coordinated posts) has the top hashtags #StopRussia' (156 posts), #StopPutin (154 posts). We see that the vast majority of these posts were aimed at a French audience, especially Emmanuel Macron, with posts such as "@EmmanuelMacron Ban Russia from SWIFT!..." (over 92% of the posts in each cluster mention "Macron" or "macron"). These posts are broadly requesting aid to stop Russia's invasion of Ukraine. In contrast, the largest cluster (20 accounts, 322 posts) hijacks terms associated with the French election to advocate for regime change in Uganda; it too has higher hazard model confidence values (0.46, Mann-Whitney U test p-value < 0.001). We show embeddings of posts from these clusters are shown in the Appendix Fig. S4b, where we see the clusters related to Russia and Putin posts have overlapping embedding distributions, suggesting the topic themes are similar (Grootendorst 2022), in agreement with the top hashtags observed. The Uganda-themed cluster, meanwhile, shows a distinct distribution.

Hazards Over Time

The model also demonstrates several surprising trends in the way hazards vary over time.

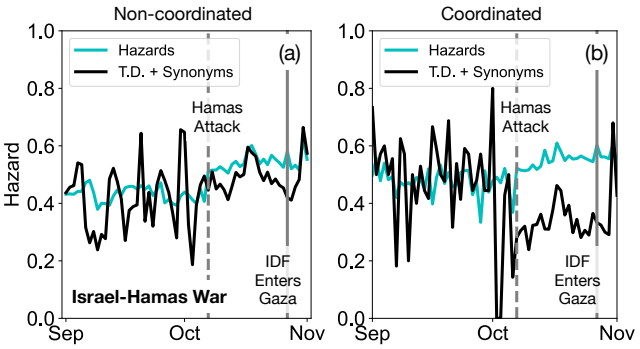

Figure 4: Hazards and threats over time for the Israel-Hamas war dataset. The plots show mean hazard confidences each day as well as the overall mean proportion of posts with at least one word from the Threat Dictionary (Choi et al. 2022) + Synonyms baseline for (a) authentic accounts and (b) inauthentic coordinated accounts. Vertical lines correspond to the Hamas attack on October 7th, 2023, and the Israel Defense Forces (IDF) entering Gaza on October 27th.

We plot hazard confidence in posts over time in Fig. 4 for the Israel-Hamas war dataset. Just after the October 7th attack, there is a sudden increase in mean hazard confidence among authentic accounts as they discuss the attack on Israeli civilians. There is no obvious increase in hazard discussions among coordinated accounts, although there is a decrease in threat words. However, there is a spike in the number of posts by coordinated accounts, shown in Appendix Fig. S1. We also see only a minor change in the proportion of posts that contain threat words (Choi et al. 2022) after the October 7th attack. Our hazard model can therefore constitute a distinct and potentially more accurate

indicator of major hazard events compared to the previous state-of-the-art methods. In a separate analysis, coordinated account posts appear to show elation (promotion of Hamas, Appendix Fig. S5, and positive emotions, Appendix Figs. S6 & S7), thus the absence of a significant change in the frequency of hazards just after October 7th occurs despite the accounts mentioning the attacks, rather than because the accounts ignore the event.

Appendix Figs. S9 & S10, meanwhile, shows hazards over time for both coordinated and authentic accounts in the 2022 French election dataset. We see notable dips during the election rounds, possibly because the vast majority of posts are positive, with posts related to "vote". We show in Appendix Fig. S8 that, with the exception of mid-May, we also see a sharp drop in coordinated account posts just after each election round. This could reflect the lower potential payoffs of information operations at these time points.

Discussion

Information concerning hazards is particularly potent, as it will frequently both garner more attention, and be more likely to be believed, than other types of information. We can therefore expect that information concerning hazards will be especially impactful in social media arenas – and, correspondingly, that this type of information will be deployed as part of calculated, coordinated social media operations intended to influence public opinion and achieve strategic objectives. Although existing approaches are effective at identifying related phenomena, such as negative affective content or terms associated with threat, none suffice for the task of pinpointing hazard information. Addressing this gap, we have developed a transformer-based model to detect hazard information in social media posts. Our model outpaces simple word-based proxies, such as the Threat Dictionary (Choi et al. 2022), and matches sophisticated LLM-based approaches with far higher throughput.

To demonstrate the model's utility, we applied it to large samples of X posts regarding, respectively, the 2023 Israel-Hamas war and the 2022 French election, both of which constitute critical recent geopolitical events, and each of which is distinct in the languages used, thus illustrating strengths of our model's ability to analyze multilingual social media data. Partitioning the data by accounts that are coordinated (and thus are likely part of information operations) and not coordinated (and thus are likely authentic accounts), we utilize our model to illuminate how hazard information is deployed in an information warfare environment.

In the Israel-Hamas war dataset, we find that coordinated accounts focus on hazards facing Gazans over hazards facing Israelis, even after an attack on October 7 that killed almost 1,200 Israelis (Blinken 2024). In the 2022 French election dataset, we find that a substantial proportion of posts are pro-Ukraine and anti-Russian, including posts in both English and French that condemn the Bucha massacre and the Russian invasion.

Overall, our analysis reveals that coordinated accounts on X often support a weaker group in a conflict, whether Hamas (opposing Israel) or Ukraine (opposing Russia). Moreover, these accounts often mention hazards impacting civilians, possibly in an attempt to evoke sympathy and enlist foreign support for their cause.

Limitations

Although we are confident that identifying hazards in text has myriad critical applications, we caution that our model's performance could still be improved. Both the subjectivity of text annotations (Davani, Díaz, and Prabhakaran 2022) and a relatively small number of annotations reduce the accuracy of many text indicator models (Chochlakis et al. 2023; Brady et al. 2021), a problem likely also present in hazard detection. Finally, while we implemented several safeguards to maximize the validity of our crowdsourced annotations, we cannot guarantee that the annotations were from organic users. Furthermore, while the coordinated accounts are suspicious, and their behavior is suggestive of an information operation, we cannot guarantee the true intention of coordinated users nor whether they are inauthentic users. There is an inevitable gray area between users who happen to post geopolitical content, even simultaneously, and users who try to influence geopolitical events.

These limitations indicate the need to iteratively improve upon our hazard detection model. These improvements can include more multilingual human-annotated data, possibly augmented with annotations by LLMs, to increase the generalizability of a model. We can also develop a multi-modal (text, images, and video) model to detect hazard information, to capture hazards in, for example, visual memes or to illuminate how traditional television media portray hazards, or use hazard information for their own editorial goals. Finally, we need to find ways to better distinguish information operations of, e.g., state actors, from other types of users so that we can better understand the intentions of these actors, especially in how they utilize hazards in text. Analysis of these data can include analyzing text clusters akin to BERTopic (Grootendorst 2022). This could uncover patterns in coordination beyond supervised labels.

Conclusion

Humans have a tendency to respond to, and share, information about hazardous events. In this work, we develop AI tools with which it is possible to observe the implications or manifestations of these tendencies at scale in online social media. Our findings are as follows. (1) We develop an openly shared tool to accurately detect information concerning hazards at scale. (2) We use this tool to reveal how this information associates with, but goes beyond, negative emotions and sentiments, as well as threat words. (3) We apply this tool to X posts about two geopolitical events, finding that information operations supporting weaker parties in conflicts often mention harms directed at civilians, potentially in an attempt to evoke sympathy and attract support for their cause.

While these findings demonstrate the utility of detecting information concerning hazards, we view our work as merely the first step, as we envision myriad applications of our model and subsequent improvements thereon, both in CS research, and in a wide variety of related fields. Given that rapidly diffusing and highly motivating hazard informa-

tion has the potential to enhance large-scale commitment to collective goals, it is vital that investigators shed light on the ways that such information is deployed in online contexts, as the same features of human psychology that present an opportunity for promoting broad cooperation toward the common good also constitute a vulnerability that can be exploited by those seeking to manipulate audiences for their own purposes.

Ethical Statement

This work was approved by an IRB. To improve account privacy, we removed all personally identifiable information from the data, such as post IDs and account IDs, prior to our analysis. Although the hazard model performs well at scale, it is still possible for the model to make mistakes. Therefore, care must be taken when interpreting the model output, including when applying it to detect whether individual accounts are sharing information concerning hazards. The confidence of the model does not guarantee that individual accounts or posts are sharing hazard content, especially given that sarcasm and in-group language constitute a challenge for current AI models. However, much like a sentiment- or emotion-detection model, so long as researchers are aware of these limitations, the harm to society is minimal; therefore, we share this model widely via the link presented at the top of the paper. We believe that the utility to society of the hazard detection model far outweighs any harmful societal impact.

Acknowledgments

This work was supported in part by NSF under award 2331722 and DARPA under contract HR001121C0168. We would also like to thank Daniel Penn, Chloe Keshishian, and Anika Scott for their help annotating text.

References

Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.; Davis, A.; Dean, J.; Devin, M.; et al. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Software available from tensorflow.org.

Acheampong, F. A.; Nunoo-Mensah, H.; and Chen, W. 2021. Transformer models for text-based emotion detection: a review of BERT-based approaches. *Artificial Intelligence Review*, 54(8): 5789–5829.

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Alhuzali, H.; and Ananiadou, S. 2021. SpanEmo: Casting multi-label emotion classification as span-prediction. *arXiv preprint arXiv:2101.10038*.

Alshehri, A.; Abdul-Mageed, M.; et al. 2020. Understanding and detecting dangerous speech in social media. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, 40–47.

authors, T. G. 2016. GPyOpt: A Bayesian Optimization framework in Python. http://github.com/SheffieldML/GPyOpt.

Badawy, A.; Ferrara, E.; and Lerman, K. 2018. Analyzing the digital traces of political manipulation: The 2016 Russian interference Twitter campaign. In *ASONAM*, 258–265. IEEE.

Bago, B.; Rosenzweig, L.; Berinsky, A. J.; and Rand, D. W. 2021. Emotion may predict susceptibility to fake news but emotion regulation does not help. *IAST working paper*.

Baumeister, R. F.; Bratslavsky, E.; Finkenauer, C.; and Vohs, K. D. 2001. Bad is stronger than good. *Review of general psychology*, 5(4): 323–370.

Bebbington, K.; MacLeod, C.; Ellison, T. M.; and Fay, N. 2017. The sky is falling: evidence of a negativity bias in the social transmission of information. *Evolution and Human Behavior*, 38(1): 92–101.

Bird, S.; and Loper, E. 2004. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, 214–217. Barcelona, Spain: Association for Computational Linguistics.

Blaine, T.; and Boyer, P. 2018. Origins of sinister rumors: A preference for threat-related material in the supply and demand of information. *Evolution and Human Behavior*, 39(1): 67–75.

Blinken, A. J. 2024. Anniversary of October 7th Attack. *Press Statement*, https://www.state.gov/anniversary--of--october--7th--attack/.

Brady, W. J.; McLoughlin, K.; Doan, T. N.; and Crockett, M. J. 2021. How social learning amplifies moral outrage expression in online social networks. *Science Advances*, 7(33): eabe5641.

Brady, W. J.; Wills, J. A.; Jost, J. T.; Tucker, J. A.; and Van Bavel, J. J. 2017. Emotion shapes the diffusion of moralized content in social networks. *PNAS*, 114(28): 7313–7318.

Brueckl, M.; and Heuer, F. 2022. *irrNA: Coefficients of Interrater Reliability – Generalized for Randomly Incomplete Datasets*. R package version 0.2.3.

Burghardt, K.; Hogg, T.; and Lerman, K. 2018. Quantifying the impact of cognitive biases in question-answering systems. In *CSCW*, volume 12, 568–571.

Burghardt, K.; Rao, A.; Guo, S.; He, Z.; Chochlakis, G.; Sabyasachee, B.; Rojecki, A.; Narayanan, S.; and Lerman, K. 2023. Socio-Linguistic Characteristics of Coordinated Inauthentic Accounts. *arXiv preprint arXiv:2305.11867*.

Cao, Y.; Hong, S.; Li, X.; Ying, J.; Ma, Y.; Liang, H.; Liu, Y.; Yao, Z.; Wang, X.; Huang, D.; et al. 2025. Toward generalizable evaluation in the llm era: A survey beyond benchmarks. *arXiv preprint arXiv:2504.18838*.

Chen, T.; and Guestrin, C. 2016. Xgboost: A scalable tree boosting system. In *SIGKDD*, 785–794.

Chochlakis, G.; Mahajan, G.; Baruah, S.; Burghardt, K.; Lerman, K.; and Narayanan, S. 2023. Leveraging Label Correlations in a Multi-Label Setting: a Case Study in Emotion. In *ICASSP*, 1–5.

Choi, V. K.; Shrestha, S.; Pan, X.; and Gelfand, M. J. 2022. When danger strikes: A linguistic tool for tracking America's collective response to threats. *PNAS*, 119(4): e2113891119.

Davani, A. M.; Díaz, M.; and Prabhakaran, V. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10: 92–110.

Dettmers, T.; Pagnoni, A.; Holtzman, A.; and Zettlemoyer, L. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36: 10088–10115.

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

Dey, P.; Luceri, L.; and Ferrara, E. 2024. Coordinated activity modulates the behavior and emotions of organic users: A case study on tweets about the Gaza conflict. In *Companion Proceedings of the ACM Web Conference 2024*, 682–685.

Dey, R.; and Salem, F. M. 2017. Gate-variants of gated recurrent unit (GRU) neural networks. In *2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS)*, 1597–1600. IEEE.

du Boisberranger, J. 2024a. RandomForestClassifier. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

du Boisberranger, J. 2024b. SVC, https://scikit-learn.org/dev/modules/generated/sklearn.svm.SVC.html.

Duppada, V.; Jain, R.; and Hiray, S. 2018. Seernet at semeval-2018 task 1: Domain adaptation for affect in tweets. *arXiv preprint arXiv:1804.06137*.

Ecker, U. K.; Lewandowsky, S.; Cook, J.; Schmid, P.; Fazio, L. K.; Brashier, N.; Kendeou, P.; Vraga, E. K.; and Amazeen, M. A. 2022. The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, 1(1): 13–29.

Fay, N.; Walker, B.; Kashima, Y.; and Perfors, A. 2021. Socially situated transmission: The bias to transmit negative information is moderated by the social context. *Cognitive Science*, 45(9): e13033.

Felbo, B.; Mislove, A.; Søgaard, A.; Rahwan, I.; and Lehmann, S. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. *arXiv preprint arXiv:1708.00524*.

Ferrara, E.; and Yang, Z. 2015. Quantifying the effect of sentiment on information diffusion in social media. *PeerJ Computer Science*, 1: e26.

Fessler, D. 2019. Believing chicken little: Evolutionary perspectives on credulity and danger. *DRUMS: Distortions, rumours, untruths, misinformation & smears*, 17–36.

Fessler, D. M. T.; Pisor, A. C.; and Holbrook, C. 2017. Political Orientation Predicts Credulity Regarding Putative Hazards. *Psychological Science*, 28(5): 651–660. PMID: 28362568.

Fessler, D. M. T.; Pisor, A. C.; and Navarrete, C. D. 2014. Negatively-Biased Credulity and the Cultural Evolution of Beliefs. *PLOS ONE*, 9(4): 1–8.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/.

Forgas, J. P. 2019. On the role of affect in gullibility: Can positive mood increase, and negative mood reduce credulity? *The social psychology of gullibility*, 179–197.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; III, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Commun. ACM*, 64(12): 86–92.

Gerard, P.; Weninger, T.; and Lerman, K. 2025. Fear and Loathing on the Frontline: Decoding the Language of Othering by Russia-Ukraine War Bloggers. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, 615–635.

Grootendorst, M. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Hama, K.; Otsuka, A.; and Ishii, R. 2024. Emotion Recognition in Conversation with Multi-step Prompting Using Large Language Model. In Coman, A.; and Vasilache, S., eds., *Social Computing and Social Media*, 338–346. Cham: Springer Nature Switzerland. ISBN 978-3-031-61281-7.

Head, T.; MechCoder; Louppe, G.; Shcherbatyi, I.; fcharras; Vinícius, Z.; cmmalone; Schröder, C.; nel215; Campos, N.; Young, T.; Cereda, S.; Fan, T.; Schwabedal, J.; Hvass-Labs; Pak, M.; SoManyUsernamesTaken; Callaway, F.; Estève, L.; Besson, L.; Landwehr, P. M.; Komarov, P.; Cherti, M.; Shi, K. K.; Pfannschmidt, K.; Linzberger, F.; Cauet, C.; Gut, A.; Mueller, A.; and Fabisch, A. 2024. scikit-optimize: Sequential model-based optimization in Python.

Hearst, M. A.; Dumais, S. T.; Osuna, E.; Platt, J.; and Scholkopf, B. 1998. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4): 18–28.

Ho, T. K. 1995. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, 278–282. IEEE.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation*, 9(8): 1735–1780.

Hong, X.; Gong, Y.; Sethu, V.; and Dang, T. 2025. AER-LLM: Ambiguity-aware Emotion Recognition Leveraging Large Language Models. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Hutto, C.; and Gilbert, E. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *CSCW*, volume 8, 216–225.

Jamil, M. L.; Pais, S.; and Cordeiro, J. 2022. Detection of dangerous events on social media: a critical review. *Social Network Analysis and Mining*, 12(1): 154.

Kartik, N.; Ottaviani, M.; and Squintani, F. 2007. Credulity, lies, and costly talk. *Journal of Economic Theory*, 134(1): 93–116.

Lilienfeld, S. O.; and Latzman, R. D. 2014. Threat bias, not negativity bias, underpins differences in political ideology. *Behavioral & Brain Sciences*, 37(3): 318.

Little, A. T. 2017. Propaganda and credulity. *GEB*, 102: 224–232.

Liu, X.; Ji, K.; Fu, Y.; Tam, W. L.; Du, Z.; Yang, Z.; and Tang, J. 2021. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*.

Luceri, L.; Pantè, V.; Burghardt, K.; and Ferrara, E. 2024. Unmasking the web of deceit: Uncovering coordinated activity to expose information operations on Twitter. In *Proceedings of the ACM Web Conference 2024*, 2530–2541.

Martel, C.; Pennycook, G.; and Rand, D. G. 2020. Reliance on emotion promotes belief in fake news. *CR:PI*, 5(1): 47.

Mazza, M.; Cresci, S.; Avvenuti, M.; Quattrociocchi, W.; and Tesconi, M. 2019. Rtbust: Exploiting temporal patterns for botnet detection on twitter. In *Proceedings of the 10th ACM conference on web science*, 183–192.

McInnes, L.; Healy, J.; and Melville, J. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Mendes, G. A.; and Martins, B. 2023. Quantifying valence and arousal in text with multilingual pre-trained transformers. In *ECIR*, 84–100. Springer.

Mohammad, S.; Bravo-Marquez, F.; Salameh, M.; and Kiritchenko, S. 2018. SemEval-2018 task 1: Affect in tweets. In *SemEval*, 1–17.

Öhman, A.; and Mineka, S. 2001. Fears, phobias, and preparedness: toward an evolved module of fear and fear learning. *Psychological review*, 108(3): 483.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine Learning in Python. *JMRL*, 12: 2825–2830.

Peng, L.; Zhang, Z.; Pang, T.; Han, J.; Zhao, H.; Chen, H.; and Schuller, B. W. 2024. Customising General Large Language Models for Specialised Emotion Recognition Tasks. In *ICASSP*, 11326–11330. IEEE.

Pennebaker, J. W.; Francis, M. E.; and Booth, R. J. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001): 2001.

Phillips, S. C.; Wang, S. Y. N.; Carley, K. M.; Rand, D. G.; and Pennycook, G. 2025. Emotional language reduces belief in false claims. *Judgment and Decision Making*, 20: e43.

Reimers, N.; and Gurevych, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Rozin, P.; and Royzman, E. B. 2001. Negativity bias, negativity dominance, and contagion. *PSPR*, 5(4): 296–320.

Sachdeva, P.; Barreto, R.; Bacon, G.; Sahn, A.; von Vacano, C.; and Kennedy, C. 2022. The Measuring Hate Speech Corpus: Leveraging Rasch Measurement Theory for Data Perspectivism. In Abercrombie, G.; Basile, V.; Tonelli, S.; Rieser, V.; and Uma, A., eds., *NLPerspectives*, 83–94. Marseille, France: ELRA.

Saha, P.; Garimella, K.; Kalyan, N. K.; Pandey, S. K.; Meher, P. M.; Mathew, B.; and Mukherjee, A. 2023. On the rise of fear speech in online social media. *PNAS*, 120(11): e2212270120.

Samore, T.; Fessler, D. M. T.; Holbrook, C.; and Sparks, A. M. 2018. Electoral fortunes reverse, mindsets do not. *PLOS ONE*, 13(12): 1–15.

Soroka, S.; Fournier, P.; and Nir, L. 2019. Cross-national evidence of a negativity bias in psychophysiological reactions to news. *PNAS*, 116(38): 18888–18892.

Stone, P. J.; Dunphy, D. C.; and Smith, M. S. 1966. *The general inquirer: A computer approach to content analysis.* MIT press.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *NeurIPS*, 30.

Veselovsky, V.; Ribeiro, M. H.; and West, R. 2023. Artificial artificial artificial intelligence: Crowd workers widely use large language models for text production tasks. *arXiv preprint arXiv:2306.07899*.

Vosoughi, S.; Roy, D.; and Aral, S. 2018. The spread of true and false news online. *Science*, 359(6380): 1146–1151.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 35: 24824–24837.

Youngblood, M.; Stubbersfield, J. M.; Morin, O.; Glassman, R.; and Acerbi, A. 2023. Negativity bias in the spread of voter fraud conspiracy theory tweets during the 2020 US election. *Humanit. soc. sci.*, 10(1): 573.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, see Methods.
   (e) Did you describe the limitations of your work? Yes, see Limitations.
   (f) Did you discuss any potential negative societal impacts of your work? Yes, see Ethical Considerations.

(g) Did you discuss any potential misuse of your work? Yes, see Ethical Considerations.

(h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, see Ethical considerations, as well as documentation/code in the code repository, discussions of anonymization in the Research Methods section.

(i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes.

2. Additionally, if your study involves hypotheses testing...

(a) Did you clearly state the assumptions underlying all theoretical results? NA

(b) Have you provided justifications for all theoretical results? NA

(c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA

(d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA

(e) Did you address potential biases or limitations in your theoretical framework? NA

(f) Have you related your theoretical results to the existing literature in social science? NA

(g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA

3. Additionally, if you are including theoretical proofs...

(a) Did you state the full set of assumptions of all theoretical results? NA

(b) Did you include complete proofs of all theoretical results? NA

4. Additionally, if you ran machine learning experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes, see the Introduction for data and code.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes, and further details are within the attached code.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes, see subsection Model Training in the Research Methods section.

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes, see Methods.

(f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes, see Ethical Considerations.

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...

(a) If your work uses existing assets, did you cite the creators? Yes

(b) Did you mention the license of the assets? Yes, the code is MIT licensed, as described in the repository link.

(c) Did you include any new assets in the supplemental material or as a URL? Yes, we share code and data in our anonymized URL.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes, data was gathered without subject's consent as data are publicly avaialble and anonymized to remove any PII.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, no data contains PII.

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? Yes, see Research Methods, page 4.

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? Yes, see the code repository link.

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? Yes, see repository link.

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? Yes, see Ethical Considerations.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? Yes, see Methods.

(d) Did you discuss how data is stored, shared, and deidentified? Yes, see Research Methods page 4.

Appendix

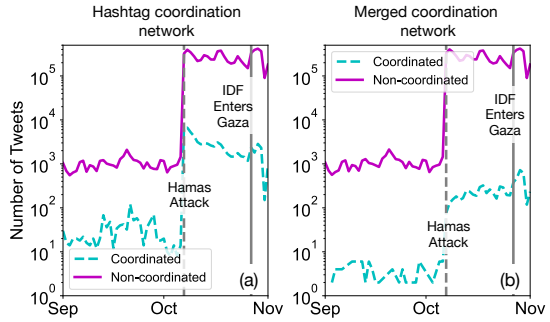

Figure S1: Post frequency over time for the Israel-Hamas dataset. Authentic and inauthentic coordinated accounts based on (a) hashtags (main text), and (b) merged similarity networks (Luceri et al. 2024). Vertical lines correspond to the Hamas attack on October 7th, 2023, and the Israel Defense Force (IDF) entering Gaza on October 27th.

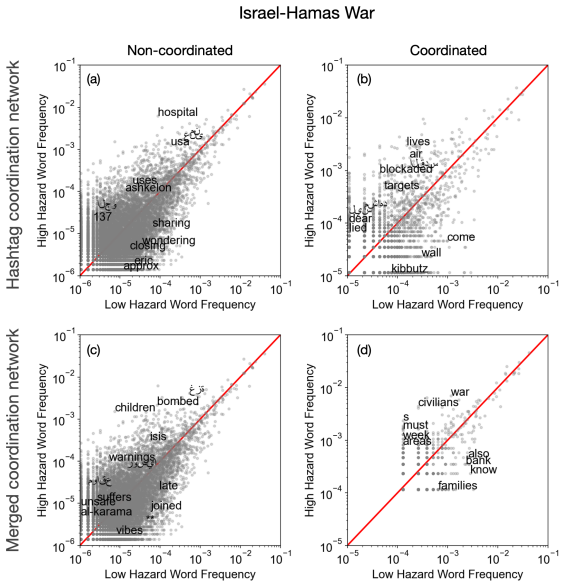

Figure S2: Words associated with high-confidence and low-confidence hazard posts among coordinated and authentic accounts within the Israel-Hamas war dataset. (a–b) Hashtag-based coordination networks, (c–d) merged coordination networks (Luceri et al. 2024).

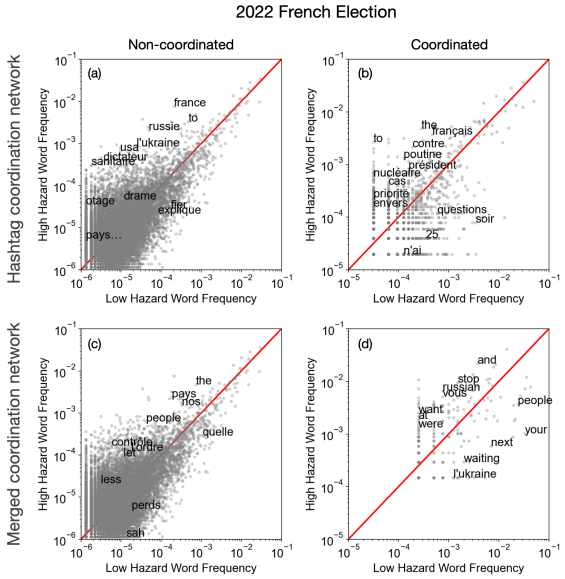

Figure S3: Words associated with high-confidence and low-confidence hazard posts among coordinated and authentic accounts within the 2022 French Election dataset. (a–b) Hashtag-based coordination networks, (c–d) merged coordination networks (Luceri et al. 2024).

Dataset	Text	Ground truth label
Random posts	Aircraft crashes into residential building in Russian city near Crimea, killing at least three https://xxxx	Hazard
Random posts	@XXXX @YYYY But he is a liar and was thrown out because he is a liar, we know you love him but really you have to be aware that he is a Liar x	No hazard
2023 Israel-Hamas War	@XX Just surreal! Footage of Palestinian Hamas terrorists who infiltrated into Israel from Gaza, firing at residents in Sderot from an SUV.	Hazard
2023 Israel-Hamas War	@XX Saudi Arabia's NEW official map scraps Israel and only names Palestine	No hazard
2022 French Election	[translated from French] @XX Macronia panics and talks about a Republican front against the left. No hesitation, let us block to their Republic the possessors, social inequalities, that of repression and police violence, for a front of anti-liberal and even anti-capitalist struggles.	Hazard
2022 French Election	[translated from French] Among the breakers of #1May2022 are the sbores of #Macron such as Roland Branleur alias @tarzoondtc troublemaker a seasoned disciple of Benala dt the objective is to decredivize the movements of struggle against the macroistic oligarchy. We wish him good health	No hazard
Dataset	**Text**	**Hazard model confidence**
2023 Israel-Hamas War	RT @XX We'll be streaming here on Monday from 1pm ET, don't miss it	0.05
2023 Israel-Hamas War	@XX Israeli airstrikes flattened mosques over the heads of worshipers. At least 2 hospitals and 2 centers run by Palestine Red Crescent Society, have been hit. So have two schools run by the U.N. agency Israel fighter jets/artillery have struck targets in Gaza frequently over years	0.92
2022 French Election	[translated from French] Is there any hope that you will have 500 referrals by Monday?	0.03
2022 French Election	@XX Putin is a murderer more dangerous than Osama bin Laden. Threats the world with a nuclear bomb. He, like no one else, can just press the red button. Gotta get him dead or alive. His immediate surroundings have the most opportunities for this. Otherwise he'll kill us.	0.94

Table S1: Hazard ground truth labels and model prediction examples.

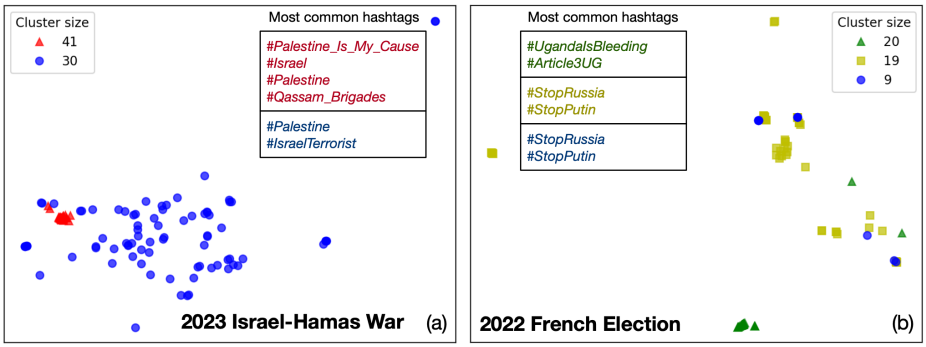

Figure S4: Embeddings text from clusters shown in in Fig. 3. (a) The 2023 Israel-Hamas war dataset, and (b) the 2022 French election dataset. Plots are colored by the cluster size, while the most popular hashtags' fonts are colored based on the cluster color. We subsample 5000 posts from each set of coordinated accounts found from the hashtag-based coordination networks, then embed the text using `distiluse-base-multilingual-cased-v2` (the same text embedding model used when detecting hazards). These embeddings are then compressed to two dimensions using UMAP (McInnes, Healy, and Melville 2018). For clarity, we only show the posts associated with the posts shown in Fig. 3 (121 and 147 for the respective subfigures).

Rank	Highest hazard word	Highest hazard ratio	Lowest hazard word	Lowest hazard ratio
1	mask	257.513	friday	0.008
2	residential	218.82	historical	0.01
3	babies	206.96	bible	0.01
4	children	205.286	jesus	0.012
5	buildings	170.045	davido	0.013
6	[translated from Arabic] children	169.452	roman	0.015
7	taxes	155.664	episode	0.016
8	[translated from Arabic] hospital	149.883	labor	0.017
9	[translated from Arabic] hospital	132.76	messi	0.019
10	woodward	132.092	rally	0.019

Table S2: Words associated with high and low-hazard posts in the full Israel-Hamas war dataset. Words have a minimum frequency of 10^{-6} in each set of posts.

Rank	Highest hazard word	Highest hazard ratio	Lowest hazard word	Lowest hazard ratio
1	children	287.486	parodique	0.011
2	war	272.263	18h	0.015
3	ukrainian	180.407	film	0.018
4	policiers	176.809	20h	0.021
5	sanitaire	172.954	netflix	0.023
6	régime	169.956	youtube	0.025
7	weapons	169.202	épisode	0.025
8	dictature	166.016	impatience	0.025
9	dangereux	146.117	émission	0.027
10	nuclear	140.831	rendez-vous	0.027

Table S3: Words associated with high and low-hazard posts in the full 2022 French Election dataset. Words have a minimum frequency of 10^{-6} in each set of posts.

Rank	Highest hazard word	Highest hazard ratio	Lowest hazard word	Lowest hazard ratio
1	openrafahcrossing	248.099	london	0.004
2	christanhospitalingaza	163.98	mia	0.005
3	communications	95.3	khalifa	0.005
4	warplanes	88.379	•	0.011
5	knew	83.055	border	0.015
6	green	81.99	received	0.016
7	starlink	80.925	friend	0.018
8	aircraft	71.342	manipulate	0.018
9	building	68.147	[translated from Arabic] Israel	0.019
10	thatś	67.881	show	0.021

Table S4: Words associated with high and low-hazard posts among coordinated accounts in the Israel-Hamas war dataset based on hashtag coordination networks. Words have a minimum frequency of 10^{-5} in each set of posts.

Rank	Highest hazard word	Highest hazard ratio	Lowest hazard word	Lowest hazard ratio
1	s	18.137	history	0.036
2	civilians	17.23	know	0.082
3	hospital	16.626	silent	0.082
4	bombed	15.87	remain	0.091
5	hospitals	14.51	feel	0.091
6	airstrikes	13.603	through	0.113
7	must	12.696	between	0.113
8	press	12.696	west	0.13
9	year	12.696	said	0.13
10	killing	11.789	happening	0.13

Table S5: Words associated with high and low-hazard posts among coordinated accounts in the Israel-Hamas war dataset based on merged coordination networks (Luceri et al. 2024). Words have a minimum frequency of 10^{-5} in each set of posts.

Rank	Highest hazard word	Highest hazard ratio	Lowest hazard word	Lowest hazard ratio
1	mask	220.641	friday	0.008
2	residential	207.869	bible	0.01
3	babies	206.401	jesus	0.012
4	buildings	196.066	historical	0.012
5	children	188.24	roman	0.015
6	[translated from Arabic] children	161.627	messi	0.015
7	taxes	146.653	davido	0.016
8	[translated from Arabic] hospital	133.148	episode	0.016
9	woodward	127.276	labour	0.017
10	gazahospitalbombing	126.395	recognized	0.019

Table S6: Words associated with high and low-hazard posts among authentic accounts in the Israel-Hamas war dataset based on hashtag coordination networks. Words have a minimum frequency of 10^{-6} in each set of posts.

Rank	Highest hazard word	Highest hazard ratio	Lowest hazard word	Lowest hazard ratio
1	mask	257.447	friday	0.008
2	residential	218.541	historical	0.01
3	babies	206.61	bible	0.01
4	children	205.17	jesus	0.012
5	buildings	169.853	davido	0.013
6	[translated from Arabic] children	169.408	roman	0.015
7	taxes	155.624	episode	0.016
8	[translated from Arabic] hospital	149.844	labor	0.017
9	[translated from Arabic] hospital	132.725	rally	0.019
10	woodward	132.058	christ	0.019

Table S7: Words associated with high and low-hazard posts among authentic accounts in the Israel-Hamas war dataset based on merged coordination networks (Luceri et al. 2024). Words have a minimum frequency of 10^{-6} in each set of posts.

Rank	Highest hazard word	Highest hazard ratio	Lowest hazard word	Lowest hazard ratio
1	to	84.905	rendez-vous	0.006
2	are	63.521	jevotezemmour	0.008
3	&	54.087	meeting	0.008
4	guerre	48.742	demain	0.008
5	of	40.88	mars	0.017
6	russie	35.849	direct	0.021
7	enfants	30.817	hashtag	0.021
8	sky	27.673	pouvez	0.022
9	ukraine	25.0	soir	0.023
10	we	24.528	[Fist emoji]	0.023

Table S8: Words associated with high and low-hazard posts among coordinated accounts in the 2022 French Election dataset based on hashtag coordination networks. Words have a minimum frequency of 10^{-5} in each set of posts.

Rank	Highest hazard word	Highest hazard ratio	Lowest hazard word	Lowest hazard ratio
1	our	42.859	future	0.028
2	children	29.355	your	0.031
3	on	19.375	where	0.045
4	killed	17.026	than	0.046
5	missiles	15.265	rt	0.059
6	these	14.678	next	0.084
7	poutine	12.916	can	0.087
8	want	9.981	do	0.09
9	russian	8.807	what	0.098
10	defense	8.807	nous	0.098

Table S9: Words associated with high and low-hazard posts among coordinated accounts in the 2022 French Election dataset based on merged coordination networks (Luceri et al. 2024). Words have a minimum frequency of 10^{-5} in each set of posts.

Rank	Highest hazard word	Highest hazard ratio	Lowest hazard word	Lowest hazard ratio
1	war	270.952	parodique	0.011
2	ukrainian	178.578	18h	0.016
3	policiers	175.836	film	0.016
4	sanitaire	172.366	20h	0.021
5	régime	169.666	netflix	0.023
6	weapons	166.993	hashtag	0.024
7	dictature	165.425	youtube	0.025
8	dangereux	146.016	épisode	0.025
9	sky	139.435	impatience	0.026
10	nuclear	136.076	émission	0.027

Table S10: Words associated with high and low-hazard posts among authentic accounts in the 2022 French Election dataset based on hashtag coordination networks. Words have a minimum frequency of 10^{-6} in each set of posts.

Rank	Highest hazard word	Highest hazard ratio	Lowest hazard word	Lowest hazard ratio
1	children	270.12	parodique	0.011
2	war	266.496	18h	0.015
3	policiers	177.052	film	0.018
4	sanitaire	173.024	20h	0.021
5	régime	170.368	netflix	0.023
6	ukrainian	168.139	youtube	0.025
7	dictature	166.34	épisode	0.025
8	weapons	162.792	impatience	0.025
9	dangereux	146.397	émission	0.027
10	criminel	128.89	rendez-vous	0.027

Table S11: Words associated with high and low-hazard posts among authentic accounts in the 2022 French Election dataset based on merged coordination networks (Luceri et al. 2024). Words have a minimum frequency of 10^{-6} in each set of posts.

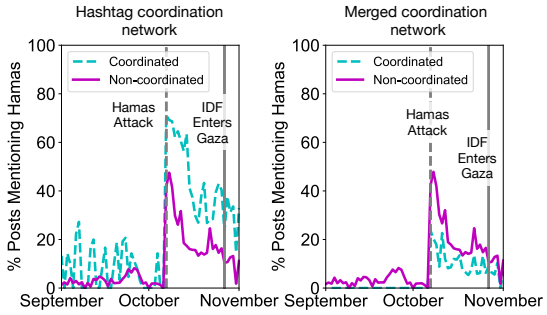

Figure S5: Percent of posts mentioning Hamas over time for (a) authentic, and (b) inauthentic coordinated accounts. Vertical lines correspond to the Hamas attack on October 7th, 2023, and the Israel Defense Force (IDF) entering Gaza on October 27th.

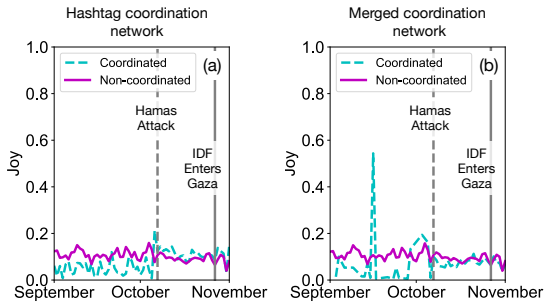

Figure S6: Mean joy emotion confidence over time for authentic and inauthentic coordinated accounts using Demux (Chochlakis et al. 2023). (a) Hashtag-based coordination network and (b) merged coordination network (Luceri et al. 2024). Vertical lines correspond to the Hamas attack on October 7th, 2023, and the Israel Defense Force (IDF) entering Gaza on October 27th.

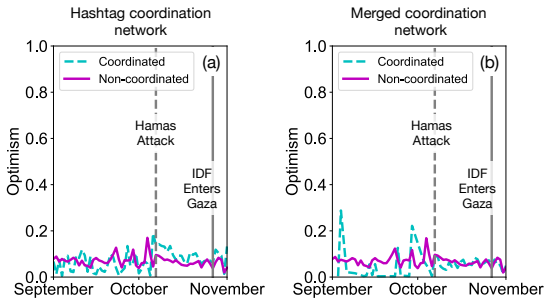

Figure S7: Mean optimism emotion confidence over time for authentic and inauthentic coordinated accounts using Demux (Chochlakis et al. 2023). (a) Hashtag-based coordination network and (b) merged coordination network (Luceri et al. 2024). Vertical lines correspond to the Hamas attack on October 7th, 2023, and the Israel Defense Force (IDF) entering Gaza on October 27th.

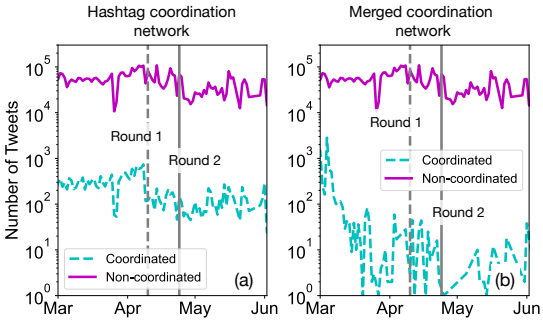

Figure S8: Post frequency over time for the 2022 French Election among authentic and inauthentic coordinated accounts. (a) Hashtag-based coordination network and (b) merged coordination network (Luceri et al. 2024).

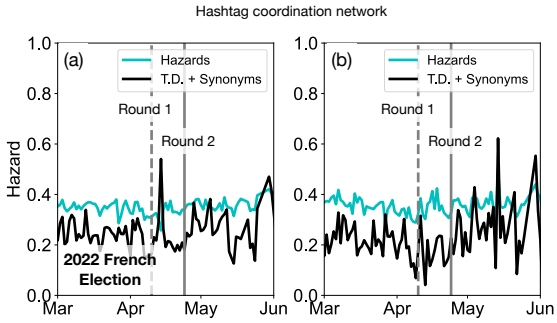

Figure S9: Hazards and threats over time for the 2022 French Election dataset. The plots show mean hazard confidences each day as well as the overall mean proportion of posts with at least one word from the Threat Dictionary (Choi et al. 2022) + Synonyms for authentic and inauthentic coordinated accounts. (a) Authentic account posts and (b) hashtag coordination network posts. Vertical lines correspond to Round 1 voting (April 10, 2022) and the Round 2 runoff (April 24, 2022).

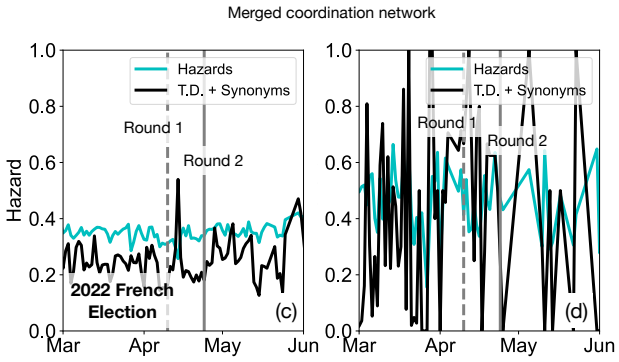

Figure S10: Hazards and threats over time for the 2022 French Election dataset. The plots show mean hazard confidences each day as well as the overall mean proportion of posts with at least one word from the Threat Dictionary (Choi et al. 2022) + Synonyms for (a) authentic accounts and (b) inauthentic coordinated accounts, where coordination is uncovered from merged coordination accounts (Luceri et al. 2024). Vertical lines correspond to Round 1 voting (April 10, 2022) and the Round 2 runoff (April 24, 2022).