Proceedings of the Eighteenth International AAAI Conference on Web and Social Media (ICWSM 2024)

Socio-Linguistic Characteristics of Coordinated Inauthentic Accounts

Keith Burghardt[1], Ashwin Rao[2*], Georgios Chochlakis[2*], Baruah Sabyasachee[2*], Siyi Guo[2*], Zihao He[2*], Andrew Rojecki[3], Shrikanth Narayanan[1, 2], Kristina Lerman[1]

[1] USC Information Sciences Institute
[2] University of Southern California
[3] University of Illinois Chicago
{keithab,ashreyas,siyiguo,zihaohe,chochlak}@isi.edu, sbaruah@usc.edu, arojecki@uic.edu, {shri,lerman}@isi.edu

Abstract

Online manipulation is a pressing concern for democracies, but the actions and strategies of coordinated inauthentic accounts, which have been used to interfere in elections, are not well understood. We analyze a five million-tweet multilingual dataset related to the 2017 French presidential election, when a major information campaign led by Russia called "#MacronLeaks" took place. We utilize heuristics to identify coordinated inauthentic accounts and detect attitudes, concerns and emotions within their tweets, collectively known as *socio-linguistic characteristics*. We find that coordinated accounts retweet other coordinated accounts far more than expected by chance, while being exceptionally active just before the second round of voting. Concurrently, socio-linguistic characteristics reveal that coordinated accounts share tweets promoting a candidate at three times the rate of non-coordinated accounts. Coordinated account tactics also varied in time to reflect news events and rounds of voting. Our analysis highlights the utility of socio-linguistic characteristics to inform researchers about tactics of coordinated accounts and how these may feed into online social manipulation.

Introduction

Social media platforms are potent vectors for manipulation (Bradshaw and Howard 2019; Badawy, Ferrara, and Lerman 2018; Kim 2018). Malicious actors use Facebook, Twitter, and other platforms to deploy inauthentic accounts that interact with and manipulate authentic users on each side of a divisive issue (Ratkiewicz et al. 2021; Kim 2018), recruit converts, incite violence, spread misinformation (Vosoughi, Roy, and Aral 2018), or undermine trust in democratic institutions (Badawy et al. 2019). Although these platforms have invested heavily to remove harmful accounts, malicious actors have adapted their strategies to evade detection and develop increasingly sophisticated influence campaigns (Sayyadiharikandeh et al. 2020). Technologies have been developed to detect, characterize, and track inauthentic account activity at scale (Ferrara 2017; Sayyadiharikandeh et al. 2020; Paper 2022), but there is a pressing need to better understand the tactics and strategies of influence campaigns that utilize inauthentic accounts through analysis of the content they promote.

In this paper, we analyze a large corpus of over 5M tweets related to the 2017 French presidential election to identify influence campaigns intended to affect the outcome of the election (Ferrara 2017). We use well-understood heuristics to identify *coordinated inauthentic accounts* (Pacheco et al. 2021) (we call these "coordinated accounts" for brevity) that may be attempting to influence election outcomes. We then create computational methods to identify attitudes, concerns, and emotions within influence campaigns. We define *attitudes* as the opinion of a user, *concerns* as the issues discussed, and *emotions* as the feelings expressed in text. Finally, we analyse how the coordinated accounts utilize these features to inform us about their tactics.

We study the French presidential election cycle, which kicked off on 10 April 2017. The first round of voting took place on 22 April 2017, with Emanuel Macron and Marine Le Pen advancing to the second round. Our motivation to analyze the 2017 election in particular is because there was a leak of French presidential candidate Emmanuel Macron's campaign emails (#MacronLeaks) on 5 May, just before the second round of voting on 7 May. #MacronLeaks leveraged a large cache of hacked documents and emails shared on WikiLeaks to discredit Macron and his party, En Marche (Ferrara 2017; Vilmer 2021), likely orchestrated by Russia (Gray 2017). It was exposed on the imageboard 4Chan and tweeted on 5 May by American alt-right activist Jack Posobiec (Gray 2017). Although the campaign ultimately failed to achieve its presumed goal (as Macron won the second round of voting) the campaign acts as an important case study of coordinated account tactics. The coordinated accounts we find are strongly over-represented in the #MacronLeaks tweets, as they were only 0.28% of all accounts but represented at least 18.7% of tweets with hashtags related to the leak within our dataset, which could represent an attempt to influence the election.

We next hypothesize a range of tactics coordinated accounts utilize through analysis of socio-linguistic characteristics. The unusual prevalence (or lack) of particular socio-linguistic characteristics within coordinated accounts compared to non-coordinated accounts helps us understand what

*These authors contributed equally.

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

coordinated accounts attempt to promote. The differences in between clusters of coordinated accounts, meanwhile, help us distinguish unique tactics that some clusters of coordinated accounts use that others do not. For example, one cluster of coordinated accounts heavily promoted concerns about national pride, international alliances, while another appeared to discuss the president of Gabon with no mention of French campaign issues. This is suggestive of multiple competing influence campaigns happening during the French election. We then show how the frequency of sociolinguistic characteristics changes over time to identify tactics, such as promoting candidates just before an election. Finally, we show how the prevalence of particular languages in each cluster hint at the different audiences for each influence campaign, such as the use of English within the pro-Marine Le Pen cluster of coordinated accounts versus French within the pro-Benoît Hamon and Francis Fillon clusters, who were round one presidential candidates. Twitter is used in France in much the same way it is used many areas of the world (e.g., for social interactions, news, political discourse, etc.), even in elections (Nooralahzadeh, Arunachalam, and Chiru 2013), thus we believe our results will generalize well outside of this election scenario.

To summarize, our contributions are the following:

- We develop novel multilingual techniques to detect socio-linguistic characteristics from tweets and make our entire pipeline publicly available.
- We use three techniques to extract coordinated networks of inauthentic accounts from Twitter users in a major election, and publicly share this code.
- We extract coordinated account behaviors and sociolinguistic characteristics.
- We apply our findings to hypothesize influence tactics.

Overall, our analysis demonstrates the feasibility of automatically identifying potential tactics used in online influence campaigns. Our code, human annotations, and example coordinated tweets are shown in the following repository: https://github.com/KeithBurghardt/Coordination.

Related Work

Political Manipulation Online manipulation is a worldwide phenomenon (cf. (Tucker et al. 2018) for a review), and can occur through a variety of ways, such as search ranking or social media trend manipulation. We specifically focus on inauthentically sharing posts that have a particular frame, a prototypical example of online manipulation (Tucker et al. 2018). This type of manipulation has long been explored on social media (Ratkiewicz et al. 2021, 2011; Kim 2018), including the Brexit vote (Howard and Kollanyi 2016), the 2016 US presidential election (Bessi and Ferrara 2016; Badawy, Ferrara, and Lerman 2018; Badawy et al. 2019; Kim 2018), the 2017 French elections (Ferrara 2017), and the 2022-2023 Russia-Ukraine war (Paper 2022). The impact of these accounts is uncertain (Bail et al. 2020), but engagement with, and attempts to manipulate, authentic users is of grave concern.

Coordinated Inauthentic Accounts Several studies also focus on the detection and behavior of coordinated accounts in social media, including on Facebook (Giglietto et al. 2020b,a), YouTube (Kirdemir, Adeliyi, and Agarwal 2022), and Twitter (Sharma et al. 2021; Nizzoli et al. 2021; Weber and Falzon 2021; Mazza, Cola, and Tesconi 2022; Cinelli et al. 2022). In contrast to bot or troll detection, coordinated account analysis focuses on detecting and analyzing accounts working in concert (Starbird 2019). Ways to uncover coordinated accounts include temporal similarities in users (Sharma et al. 2021; Weber and Falzon 2021; Schliebs et al. 2021; Pacheco et al. 2021), similarity in content (Schliebs et al. 2021), comment networks (Kirdemir, Adeliyi, and Agarwal 2022), URLs shared (Giglietto et al. 2020a), user attributes, and co-retweeting (Pacheco et al. 2021; Mazza, Cola, and Tesconi 2022).

Most of these studies analyze these coordinated campaigns within elections, although there are exceptions to this trend, such as coordinated accounts related to COVID-19 (Graham et al. 2020; Piña-García and Espinoza 2022). The goals of coordinated accounts, however, are less-studied. While previous work includes analyzing stories promoted by coordinated accounts (Ehrett et al. 2021), or stances by social bots (Chen et al. 2021), there is a lack of research on socio-linguistic characteristics expressed by coordinated accounts, and how they may feed into manipulation tactics.

Attitude Analysis Attitudes, such as voting for or against a candidate are a distinct set of tools we utilize in this paper, but have analogues in previous work. Attitudes most closely resemble stances (for a review, cf. (Küçük and Can 2020)), previously used to study misinformation (Hardalov et al. 2022), as they aim to determine the opinions users are trying to convey. Meanwhile, some attitudes, such as the belief that a candidate is corrupt, are similar to moral framing (Linvill, Warren, and Moore 2021), whereby an action is viewed as a virtue or vice, or person is viewed as virtuous or corrupt.

Concern Analysis Concerns, meanwhile, represent key topics discussed by Twitter users, and have analogues to topic modeling (Mei et al. 2007; Eisenstein, Ahmed, and Xing 2011; Jelodar et al. 2019), framing (Card et al. 2015), as well as position issues (Stokes 1963) that divide voters. Among the many possible topics, we focus on those discussed by the French presidential election (Lachat and Michel 2020).

Emotion Analysis Emotion extraction tools have perhaps the longest history, starting with the General Inquirer (Stone, Dunphy, and Smith 1966), and were iteratively improved with dictionary-based methods, such as LIWC (Pennebaker, Francis, and Booth 2001), EmoLex (Mohammad and Turney 2010), and DDR (Garten et al. 2018). Alike to dictionary-based methods, bag-of-words features have been used along side other features to build emotion recognition systems (Wang and Pal 2015; Li et al. 2015a), including sentence-level emotion predictions (Li et al. 2015b).

The most successful emotion recognition methods deploy Deep Learning (He and Xia 2018), such as those based on

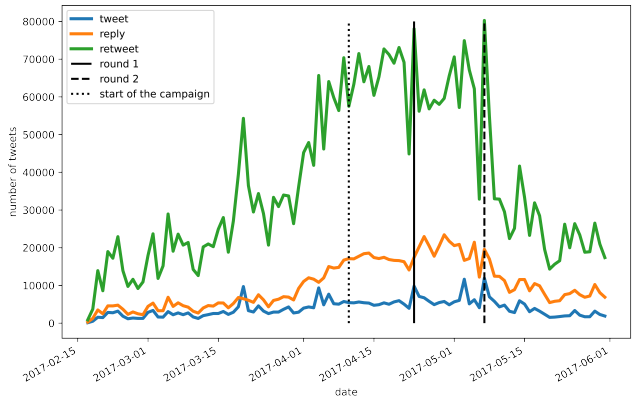

Figure 1: Daily number of tweets, retweets and replies in the 2017 French election data. Vertical lines mark important events: start of the campaign (dotted line), first round of voting (solid line), second round of voting (dashed line).

LSTMs (Hochreiter and Schmidhuber 1997; Duppada, Jain, and Hiray 2018; Yu et al. 2018), and bidirectional LSTMs (Baziotis et al. 2018). Recently, Transformers (Vaswani et al. 2017) have dominated the field. Ying et al., (2019) use the `[CLS]` token of BERT along with a shallower Convolutional Neural Network meant to learn task-specific n-gram patterns to predict emotions. Other methods use Graph Neural Networks (Xu et al. 2020), and took advantage of correlations between emotion (Alhuzali and Ananiadou 2021). In this paper, we use the current state-of-the-art algorithm, Demux (Chochlakis et al. 2023), which outperforms competing methods on the SemEval 2018 Task 1 e-c benchmark (Mohammad et al. 2018).

Methods

We present the data and the methods we use to extract sociolinguistic characteristics from tweets.

Data

We apply our methods to a corpus of 5.3M tweets about the 2017 French presidential election and automatically detected the attitudes, concerns, and emotions in each tweet. The tweets were collected by querying Twitter with a set of keywords related to the election: e.g., "election", "élection", "l'élection", "Elysee 2017", "Elysee2017", etc. (Ferrara 2017). In addition, collected tweets include those posted by accounts of presidential candidates, their parties or campaigns, such as @MLP_officiel, @EmmanuelMacron, @enmarchefr, @JLMelenchon, and @jlm_2017. The vast majority (91%) of tweets were in French, 4% were in English, and the rest were a wide variety of other languages including 3% unknown based on the Twitter API's language detection feature.

Fig. 1 shows the daily volume of messages. Online discussion geared up long before the official start of the presidential campaign (10 April 2017) with sharp peaks on the days of the first (23 April 2017) and second (7 May 2017) rounds of voting. Interest in the campaign dropped sharply thereafter, with small increases around the time of Macron's inauguration (14 May 2017). Although quote tweets were used in 2017, they are missing in our data, therefore in the rest of the paper, we analyze original tweets, replies, and retweets.

Attitude Detection

Attitudes describe what a message's author thinks and believes. In the context of an election, influence messages express attitudes that promote a candidate or party either by explicitly telling voters to vote for or against them or by using moral outrage (e.g., saying a candidate is immoral) to drive people from opposing candidates and parties. Moral framing, such as framing a candidate or party as corrupt (Linvill, Warren, and Moore 2021), is a powerful motivator strongly linked to partisan identity (Graham et al. 2018). For all attitude indicators we used NVIDIA Tesla A40 or A100 GPUs in an internal cluster.

Vote for or Against CANDIDATE or PARTY To detect the author's attitude towards the target, i.e., CANDIDATE or PARTY, we frame the problem as stance detection (cf. (Küçük and Can 2020; Hardalov et al. 2022)). The detected stance can be "in favor," "against" or neutral: e.g., if we find that a tweet is in favor of Macron, then its attitude is "vote for Macron". Here, CANDIDATE or PARTY is a wildcard that represents any of the 11 candidates in 2017 presidential election or their associated parties, including the run-off candidates Macron (party: En Marche) and Le Pen (party: Front National, later renamed Rassemblement National in 2018).

We encode each pair consisting of a text (tweet) and a target (CANDIDATE/PARTY) with a pretrained multilingual text embedding model, XLM-T (Barbieri, Anke, and Camacho-Collados 2022). This representation is then fed into a feed-forward neural network for stance classification. This intuitive method poses two challenges. First, inferring the stance entails some background knowledge about the target; second, tweets labeled with stances towards candidate in the 2017 French Election are scarce, making supervised learning difficult.

To address the first challenge, we use a stance detection model WS-BERT (He, Mokhberian, and Lerman 2022) that uses relevant Wikipedia entries for background information about the target needed to infer stance. By using XLM-T embeddings, instead of BERT used previously (He, Mokhberian, and Lerman 2022), however, our method can extend to multilingual data. To meet the second challenge, we pre-train the model on two other supervised stance detection datasets, COVID-19-Stance (Glandt et al. 2021) and P-Stance (Li, Zhao, and Caragea 2021), and then fine-tune this model on 10K human annotated tweets, described later. We apply this model to infer the stance about 11 candidates. COVID-19-Stance has tweets in a COVID-19 domain annotated with "favor" and "against" for "Anthony S. Fauci, M.D.", "Keeping Schools Closed", "Stay at Home Orders", and "Wearing a Face Mask." P-Stance has tweets in a political domain annotated with "favor" and "against" for "Biden", "Sanders" and "Trump", which lies in a political domain similar to our case.

CANDIDATE or PARTY is Moral or Immoral We operationalize moral judgment using Moral Foundations (MF) Theory (Graham et al. 2013), which proposes five dimensions of morality, each with its virtues and vices: care vs. harm, fairness vs. cheating, loyalty vs. betrayal, authority vs. subversion, and sanctity vs. degradation. We consider all the virtues to define the class "moral," and all the vices as the class "immoral."

For this model, we first pre-process all tweets by removing URLs, replacing all mentions with "@user", removing or split hashtags, converting emojis to a description, converting text to lower case removing punctuations and non-ascii text, and removing emoticons.

We then train our model first using Moral Foundations Tweet Corpus (MFTC) (Hoover and et al. 2019), which contains English language tweets annotated by the morality they express, and then fine-tune the model with 10K human annotated French tweets. For each tweet, we take majority vote as the true label. We then fine-tune a pre-trained multilingual model XLM-T (Barbieri, Anke, and Camacho-Collados 2022) with a binary prediction layer (a sigmoid activation). The model allows for multi-label prediction, because a tweet may express more than one moral judgment. We further finetune this model using 10K human annotated tweets. Although we do not have an equivalent French moral dictionary, the XLM-T multilingual embedding allows our model to transfer knowledge from English words to the majority-French dataset.

Concern Detection

Concerns are divisive issues that separate potential voters into distinct blocs, i.e., position issues (Stokes 1963). We focus on a subset of the issues salient to the 2017 French presidential election (Lachat and Michel 2020), namely Economy, Terrorism, Religion, Immigration, International Alliances. Russia Relations, National Identity, Environment, Misinformation, and Democracy. For all concern indicators we used NVIDIA Tesla A40 or A100 GPUs in an internal cluster.

To detect concerns, we fine-tune a BerTweetFr model (Guo et al. 2021) to predict concerns from 10K human annotated data and train for 3 epochs with a batch size of 8. Each concern becomes a binary label prediction task, allowing for multiple concerns to be found in each tweet.

Emotion Detection

Emotions are feelings expressed in a message. Even a short text—a tweet—can convey emotions. The emotional expression spans a range from anger and hate to joy and pride. For all emotion indicators we used NVIDIA Tesla A40 or A100 GPUs in an internal cluster.

Our emotion detection tool is based on Demux (Chochlakis et al. 2023), which is the state-of-the-art model trained on SemEval 2018 Task 1 E-c (extracting emotions from text) (Mohammad et al. 2018). Demux includes the names of emotions in the input as its first input sequence, and the actual input as the second sequence. The contextual embeddings for each emotion are used to get a confidence. Consequently, the model can predict none, one, or multiple emotions per input. We apply XLM-T (Barbieri, Anke, and Camacho-Collados 2022, 2023) to Demux to improve multilingual emotion prediction.

To simplify emotion recognition, similar emotions that often co-occur are grouped into clusters. Our approach attempts to automatically recognize these clusters: "Anger, Hate, Contempt and Disgust", "Embarrassment, Guilt, Shame and Sadness", "Admiration and Love", "Optimism and Hope", "Joy and Happiness", "Pride and National Pride", "Fear and Pessimism", "Amusement", other positive emotions, and other negative emotions. These labels combine similar emotions, and account for nuances of the French election (e.g., discussion of pride, including national pride). Amusement meanwhile is not an emotion per se, but we find is often evoked in tweets.

Using the English and Spanish tweets in SemEval 2018 Task 1 E-c for pre-training (Duppada, Jain, and Hiray 2018), we combined anger and disgust into "Anger, Hate, Contempt and Disgust"; sadness into "Embarrassment, Guilt, Shame and Sadness"; love into "Admiration and Love"; optimism into "Optimism and Hope"; joy into "Joy and Happiness"; and fear and pessimism into "Fear and Pessimism". The other labels were not pre-trained. Due to the multilingual nature of these embeddings, pre-training on non-French data does not harm the model. We then fine-tuned the model with 10K human annotations of French tweets, which have support over all the emotions.

Fine-Tuning And Evaluation Dataset

An independent Testing & Evaluation (T&E) team is used to annotate 10K French election tweets. The T&E team recruited and trained 15 annotators who were all fluent French speakers and actively followed French politics. They were given an annotation guide document written in English (shown in https://github.com/KeithBurghardt/Coordination/blob/main/annotations/README.md), describing what each attitude (called an "agenda" in the document), concern, and emotion represents. Annotators were given a small subset of these 10K tweets such that at least three annotators labeled each tweet for each socio-linguistic characteristic. These labels were all binary and a tweet could contain multiple attitudes, concerns, or emotions. The unweighted mean inter-annotator agreement, κ (Cohen 1960) is 0.51 for attitudes, 0.67 for concerns, and 0.34 for emotions, which represents fair to substantial agreement (McHugh 2012).

To evaluate the models, we reshuffle these 10K tweets and take the first 5K for training while holding out the next 5K for testing. We then compute the ROC-AUC for each attitude, concern, and emotion. This process is repeated ten times to calculate the variance of the performance metrics. The results are shown in Fig. 2. Our models generally achieve high ROC-AUC scores, which gives us confidence in their ability to detect these features.

Coordinated Inauthentic Accounts

Coordinated accounts are accounts that work together towards some broader objective while seeking to mislead people about their goals (Giglietto et al. 2020b; Pacheco et al. 2021; Cinelli et al. 2022). Such accounts could

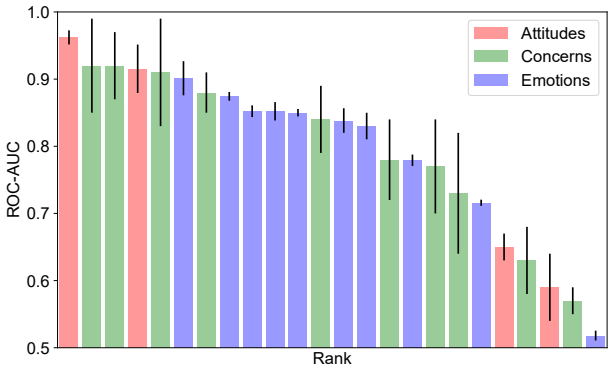

Figure 2: Evaluation of the models' predictions on a 5K subset of 2017 French Election tweets. The bars show AUC scores predicted by the models, ranked from highest to lowest ROC-AUC in held-out data: Vote for attitude, Economy and Terrorism/counterterrorism concerns, Vote against attitude, Religion concern, Embarrassment emotion, Immigration concern, Anger, Pride, Sarcasm, and Hope emotions, Environment concern, Admiration and Fear emotions, Misinformation concern, Positive-other emotion, Russia and Democracy concerns, Negative-other emotion, Immoral attitude, National identity concern, Moral attitude, International alliances concern, and the Joy emotion. Black lines indicate standard errors after bootstrapping ten times (see Methods). All results are statistically significantly above the 0.5 baseline (z-score p-value < 0.05) except for Morals (p-value$= 0.07$).

be social bots (Ferrara et al. 2016), or humans, e.g., paid trolls (Badawy, Ferrara, and Lerman 2018). Due to the Twitter terms of service that we follow, we can not check if accounts are bots as all data, including usernames, are anonymized. Moreover, even if the data were not anonymized, the high false positive rate for bot detection (Rauchfleisch and Kaiser 2020) makes insights about bots more difficult to infer. To collect networks of coordinated accounts, we identify pairs of accounts with unexpectedly similar behaviors (Nizzoli et al. 2021), namely those whose original tweets share five or more hashtags in the same order, which represents tweets that are semantically very similar. We do not claim that this method creates an exhaustive list of coordinated accounts in the dataset. However, this heuristic can detect the largest number of likely coordinated accounts compared to alternative methods (Pacheco et al. 2021), such as timing of messages, sharing user profile information, sharing of what is retweeted, and other features (Giglietto et al. 2020b). For robustness, however, we compare this method against two alternatives: retweet similarity and tweet time similarity. The former is defined as taking a TF-IDF vector of all tweets that are retweeted in the dataset. The top 0.5% of cosine similar users that have more than ten retweets in the dataset are considered coordinated. We will show that this method has drawbacks. We contrast this method with tweet time similarity. To calculate this metric, we first extract the time any tweet (original, reply, or retweet) was sent for each account that has sent more than ten tweets. We bin these tweets into 30 minute intervals, and convert the series of binned tweet times for each account into a TF-IDF vector. If the cosine similarity of these accounts is > 0.99 (this is an arbitrary cutoff; results are robust to this choice) then we consider the accounts coordinated. For all coordinated account extraction, we used Intel(R) Xeon(R) CPUs in an internal cluster.

Results

We extract socio-linguistic characteristics of tweets to study user behavior during the election cycle, how people respond to external events, and to elucidate coordinated account tactics within information campaigns.

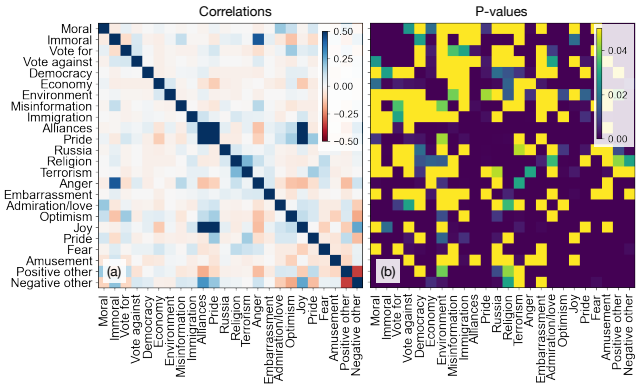

Figure 3: Spearman rank correlations between attitudes, concerns, and emotions for 10K human annotated tweets. (a) Correlations and (b) p-values of the correlations.

Correlation of Socio-Linguistic Characteristics

We first spot check the validity of the socio-linguistic characteristics by analyzing their correlations. Figure 3a shows Spearman rank correlations between all characteristics within the 10K human annotations while Fig. 3b shows the p-values of these correlations. In agreement with expectations, attitudes in support of a candidate or party ("vote for," "is moral") are correlated with each other and with positive emotions (Admiration, Optimism, Joy, Pride) and are anti-correlated with their opposed attitudes ("is immoral") and negative emotions (Anger, Embarrassment, Fear). Surprisingly, "vote against" is correlated with "vote for" possibly because there is ambiguity in whether tweets discuss voting for one candidate or against another. Positive emotions are correlated with each other as are negative emotions, in agreement with previous work (Alhuzali and Ananiadou 2021), and each type of emotion is anti-correlated with its opposite. The only exception is "amusement," which is correlated with negative emotions and anti-correlated with positive; this is consistent with the emotion representing sarcasm. We also find the "economy" concern is correlated with "immigration," "environment," and "international alliances,"

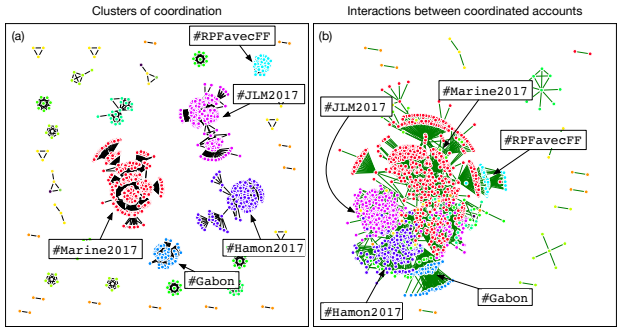

Figure 4: Coordinated networks. (a) Nodes represent Twitter accounts and links connect accounts that share at least one original tweet with the same sequence of five or more hashtags. The most popular hashtag is listed next to the five largest connected components. (b) Retweets between coordinated accounts. Cluster colors are the same in both subfigures.

while "misinformation" is correlated with "international alliance." Finally, the "national pride" concern is correlated with the emotion pride.

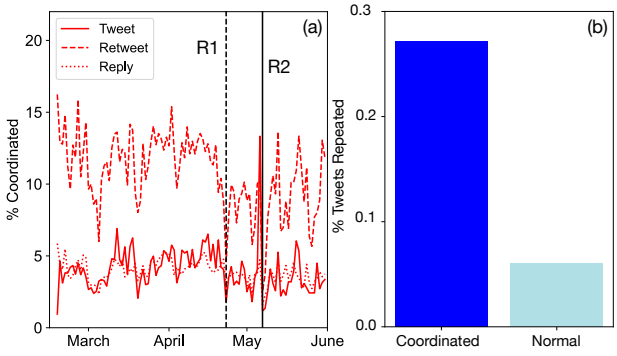

Figure 5: Coordinated activity patterns. (a) Share of original tweets (solid line), retweets (long dashed), and replies (short dashed) that are from coordinated accounts. Another common message type, quote tweets, are not found in our dataset (which was collected by a third party), and are therefore not included in this plot. (b) Share of duplicate tweets posted by accounts.

Coordinated Account Tactics

We next identify networks of coordinated accounts and analyze their behavior.

The Network of Coordinated Inauthentic Accounts Fig. 4a shows the identified network of coordinated accounts (a total of 1.6K accounts). The accounts are linked if they tweeted five or more of the same hashtags in the same order. We see several connected components, which we call coordinated account clusters.

When we analyze the text of these coordinated accounts, we find 1.6K tweets that contained any of three hashtags often representing the #MacronLeaks story: #MacronLeaks, #Bayrougate, or #Macrongate, while the total number of tweets in the dataset with these hashtags is 8.9K. Coordinated accounts were therefore responsible for at least 18.7% of these conspiracy tweets despite representing just 0.28% of all users in our dataset. In Fig. 4b, meanwhile, we show retweets between these coordinated accounts, which shows a surprising number of interactions. In total, 10.7K retweets or 33% of coordinated account content is retweeted by other coordinated accounts. These retweets are likely a tactic to promote content to each other's wider audiences. This tactic appears to be successful as there are 6.9K replies and 22K retweets of coordinated accounts by likely non-coordinated users.

We give an overview of coordinated account behavior in Fig. 5. We see in Fig. 5a that coordinated accounts are responsible for a disproportionate number of tweets. They represent only 0.28% of all accounts yet created $\sim 5-10\%$ of tweets, replies and retweets. Just before the second round of voting, original tweets from coordinated accounts became even more prominent, possibly to promote particular candidates or to discredit Macron through #Macronleaks. Finally, we notice a much larger proportion of coordinated account tweets were duplicates compared to normal users (Fig. 5b). The difference is statistically significant (Mann-Whitney U test p-value $< 10^{-10}$), and our results are robust if we remove URLs or username mentions.

When we analyze individual coordinated account clusters, we notice different presidential candidates and parties are prominent. The largest cluster (927 users) used hashtags that support Le Pen (#LePen, #Marine2017) and often promoted conspiracies about Macron (they tweeted #MacronLeaks 682 times, more than twenty times any other cluster). Other clusters supported Emmanuel Macron and Benoît Hamon (the three most frequent hashtags are #Hamon2017, #EnMarche, and #JeVoteMacron in that order; 162 accounts) or Jean-Luc Mélenchon and La France Insoumise (the two most frequent hashtags are #JLM2017, #Franceincoumise in that order; 309 accounts). The latter set of coordinated accounts also promoted hashtags such as #JulieLançon and #JURA, which are words related to the French 2017 legislative election on Jul 11 and 18th. Namely, Julie Lançon was a La France Insoumise candidate in the election within Jura's 2nd constituency. Although Lançon received only 3,323 votes in the 2017 election, at least 86% of #JulieLançon tweets in our dataset (161 out of 187) were created by coordinated accounts; she was supported by 5.4% or roughly one in twenty coordinated accounts we detected.

We also notice a surprising cluster of 57 accounts with hashtags that include #Gabon (the most popular hashtag), and unrelated hashtags in order of popularity #ZDF (the German public-service broadcaster), #10Mai2017_A_Geneve, and #i, presumably to be seen in a range of Twitter conversations unrelated to Gabon. Several times the accounts mention Gabon president, such as #BongoIsKilling (where Ali Bongo Ondimb was Gabon's president in 2017). Tweets include, "je rêve d'un Gabon Unis sans Bongo, d'un Gabon à l'abri de la peur et du besoin #SOSGABON..." which trans-

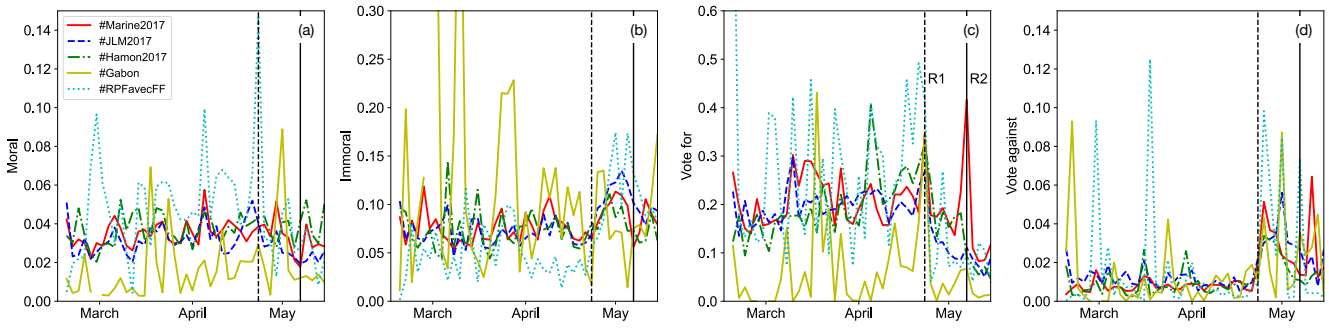

Figure 6: Attitudes over time for coordinated account clusters. Mean confidence (classifier prediction score) over time for tweeting (a) a person or group is moral, (b) a person or group is immoral, (c) voting for a candidate, or (d) voting against a candidate.

lates to "I dream of a United Gabon without Bongo, of a Gabon free from fear and need #SOSGABON".

Finally, the smallest cluster is one that promotes Francis Fillon (the three most frequent hashtags are #RPFavecFF, #RPF, and #Fillon2017 in that order; 35 accounts), where #RPF is a defunct political party, thus the account appears to want voters from a former party to vote for Fillon. Moreover, this account contains the #Grasse hashtag, which is in reference to a shooting in the town of Grasse. The diversity of hashtags and topics within each cluster suggests that multiple, sometimes competing, influence campaigns were simultaneously active during the presidential election.

Socio-Linguistic Characteristics of Coordinated Accounts For more insight into campaign tactics, we analyze how the mean confidences of socio-linguistic characteristics over time, where we plot attitudes over time in Fig. 6. Figs. 6a–b shows that discussions of moral candidates or political parties decreases slightly between rounds one and two, but immoral claims spike just before round two. We also notice that discussions of voting for a candidate peak in round one and then decrease for clusters #JLM2017, #Hamon2017, and #RPFavecFF, where the later two clusters are related to candidates who lost. Discussions about voting strongly peak in round two for #Marine2017, suggesting a strong advocacy to vote for Marine Le Pen within that account cluster. Voting against opposition candidates across all clusters, meanwhile, peaks between rounds one and two. This also parallels analyses of emotions over time (not shown), where negative emotions peak between round one and two. This agrees with analysis of 10K human annotated data, where we find anger, fear, and negative-other correlates with voting against a candidate (Spearman rank correlations$= 0.06,\ 0.03,\ 0.08$, p-values≤ 0.001).

While model confidences are used throughout the paper, we can also binarizing labels. For example, a tweet with confidence 0.8 that it contains the love/admiration emotion is then given the label 'love/admiration'. This results in virtually identical results. Namely, while the Spearman correlation between confidences and binarized labels across all 5M users aggregated at the daily level vary for each socio-linguistic characteristic, the median correlation is high at 0.85.

We next analyze time averaged socio-linguistic characteristics within coordinated account clusters in Fig. 7. This figure takes the difference in the mean tweet confidence between accounts withing a coordinated network or cluster and all ordinary (non-coordinated) accounts. All results are significant based on the Mann-Whitney U test (p-values $<$ 0.05) except: the attitude "vote against" for #RPFavecFF, the concern "alliances" for #Gabon, and the emotions "negative-other" for #RPFavecFF, and "anger" and "embarrassment" for #JML2017. There are many similarities across coordinated networks. First in Fig. 7a, coordinated accounts tweet more about the voting for candidates (vote for attitude). To put these values in perspective, if we binarize labels for each tweet, we find that 35% of all coordinated account tweets promote a candidate or party in contrast to just 8.2% among non-coordinated users. In Fig. 7b, we find that larger coordinated account clusters have lower religion, alliances, and immigration confidences, with the exception of the #Marine2017 cluster. Finally, in Fig. 7c, coordinated accounts tend to have lower amusement, embarrassment, and admiration/love confidences.

Key differences, however, abound. Most notably the #Gabon cluster's attitudes have lower voting or positive moral stances confidences but a higher immoral stance. Meanwhile, their terrorism concern confidences are higher, and economic concern confidences are lower than non-coordinated accounts. Finally, their tweets are very negative with low optimism or positive emotions. This reflects their typically off-topic and admonishing tweets about Gabon's president. The #Marine2017 cluster, meanwhile is unusual by having higher religion, national price, alliance, and immigration confidences than non-coordinated users (and most coordinated clusters). The #Marine2017 cluster therefore appears to be diving deep into divisive issues, perhaps to separate Marine from other candidates or perhaps to create wedge issues that divide the electorate.

There are a number of coordination metrics (Pacheco et al. 2021), therefore, to check the robustness of our results, we also determined coordination based on similarities of retweets (24 accounts, no accounts overlap with hashtag-

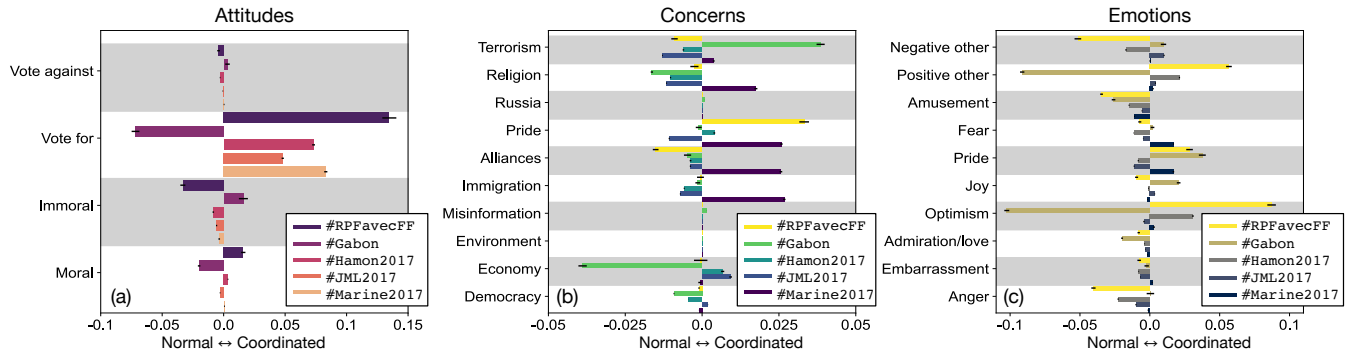

Figure 7: Socio-linguistic characteristics used by coordinated information campaigns. (a) Attitude, (b) concerns, and (c) emotions for each of the five largest coordinated account clusters (sorted from top to bottom: #RPFavecFF, #Gabon, #Hamon2017, #JML2017, and #Marine2017). The x-axis shows the difference between the mean tweet confidence of each cluster compared to non-coordinated campaign tweets. Positive values indicate coordinated account tweets whose socio-linguistic characteristic confidences are higher than non-coordinated accounts, and negative values indicate confidences that are lower. Black lines represent standard errors.

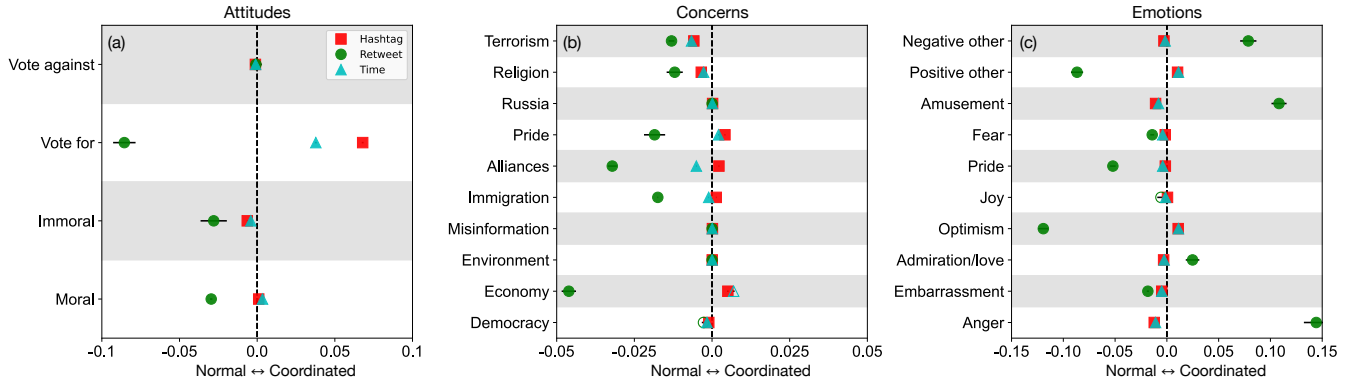

Figure 8: Socio-linguistic characteristic confidence differences between normal and coordinated networks. Coordination is defined as sharing unique sequences of hashtags in a tweet ("hashtag" - 1.6K users), similarities of retweets ("retweet" - 24 users), or similarities of tweet times ("time" - 404 users). (a) Attitudes, (b) concerns, and (c) emotions. Black lines represent standard errors. Open markers represent values not statistically significantly different from 0 (Mann-Whitney U test p-values > 0.05).

based coordinated accounts), and the timing of tweets (404 accounts, 108 overlapping with hashtag-based coordinated accounts). The results are summarized in Fig. 8, where we take the difference in the mean confidences between coordinated and non-coordinated accounts. This figure contrasts with Fig. 7 by studying the aggregate differences between coordinated and non-coordinated accounts (across all clusters) for each coordination metric, rather than measuring each cluster for one coordination metric. We find consistent behavior between hashtag and tweet time-based coordinated accounts, where about 27% of the tweet time-based set of coordinated accounts are also in the hashtag-based set of coordinated accounts. Retweet-based accounts show distinct behavior both because of the small number of (possibly non-representative) accounts and because the accounts may utilize a different set of manipulation tactics.

Not only can we capture cluster-level behavior, our analysis can also reveal differences in individual coordinated accounts, which we show in Fig. 9. Several findings are apparent in Fig. 9a. First, the #Marine2017 cluster stands out for having surprisingly few tweets in French (whose language is indicated by the Twitter API), with only 36% of tweets in French on average for each account while 48% are in English. This agrees with previous finding that a majority of tweets in the #MacronLeaks campaign were in English (Vilmer 2021). There are also many non-French accounts in the #Gabon cluster, although there is greater uniformity. Next, we demonstrate the diversity of socio-linguistic characteristics across accounts with a case study in Fig. 9b, which shows the mean vote for attitude confidence across all tweets for each coordinated account. The confidence is especially high for the #Marine2017 cluster although there is a wide variance. In contrast, the #Gabon cluster has very low vote for attitude confidence across all accounts.

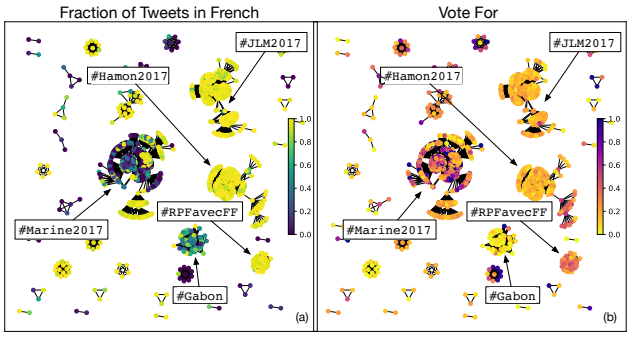

Figure 9: Socio-linguistic characteristics within information campaigns. Mean tweet confidence per user for (a) fraction of tweets in French and (b) vote for attitude.

Discussion

The results demonstrate key differences in socio-linguistic characteristics of coordinated accounts and non-coordinated accounts, which provides a nuanced understanding of coordinated accounts. First, these results show that coordinated accounts retweet a disproportionate amount, especially retweeting each other to amplify their messages, and repeat their tweets more often than ordinary users in order to amplify exposure through repeated messaging. This is a tactic useful to increase online attention (Cox and Cox 2002). Next, the socio-linguistic characteristics demonstrate that coordinated accounts attempt to push a "vote for candidate" message far more than non-coordinated accounts, especially before elections, possibly to guide potential voters to a very specific candidates. Interestingly, we also found negative emotions increased between election rounds. Negative campaigns can be effective if done correctly (Fridkin and Kenney 2004), and attacks against candidates could be more believable (Fessler, Pisor, and Navarrete 2014). Moreover, coordinated accounts selectively push particular election concerns (most notably, the #Marine2017 coordinated cluster discussed national pride, alliances, and immigration). The results also show coordinated accounts promoting small elections, such as the candidate Julie Lançon. The outlier #Gabon cluster meanwhile does not seem to advocate for a particular candidate but instead mentions prominent Twitter accounts (every single #Gabon account tweet mentions a Twitter user), including French politicians. This may be a tactic for these tweets to appear in Twitter searches for the politician's Twitter handle (an especially likely scenario during an election). This will allow a wider international audience to see these messages.

Our work highlights a number of wider implications. Namely, coordinated accounts have remarkable diversity the agendas, concerns, and emotions they share, even within closely aligned clusters. There is no exemplar markers of influence campaigns, presumably because these coordinated accounts attract a different audience. Related to this, the type of tweets vary over time; for elections this can be to promote (vote for) or attack (vote against) parties and candidates, yet coordinated account clusters often retweet each other (in agreement with a previous paper (Wang et al. 2023)). This may be a tactic boost each other's messages.

Conclusions

Our analysis of a large body of tweets related to the 2017 French election reveals psycho-social dynamics of coordinated accounts. While the coordinated accounts we identified were only 0.28% of all users, they comprised of 5-10% of all retweets and at least 18.7% of #MacronLeaks tweets, an information campaign led by Russia. Consistent with this, we also find coordinated account activity spiked just before round two (when the #MacronLeaks story first appeared). Coordinated accounts appear to have employed a range of tactics, such as repeating their messages, sharing positive content ("vote for" rather than "vote against"), sharing more positive emotions, and focusing on some voter concerns, such as national pride and the economy. That being said, we also notice a degree of diversity in coordinated account clusters, possibly because these clusters are tailored to different audiences. Overall, the results point to coordination accounts being used for social manipulation and we uncover potential tactics towards that purpose.

While our methods have given new insights into coordinated accounts, they have a number of limitations that motivate future work. Namely, we find the socio-linguistic characteristic models are imperfect. This is a limitation, which should be improved in the future. Part of this limitation is due to data imbalance, therefore more data, especially for low-support classes is critical. Next, the data is a biased sample (Morstatter et al. 2013), which limits the generalizability of our findings. A more representative sample, especially of recent elections is needed to validate these findings. In addition, the degree to which these results generalize outside France or outside of elections needs to be studied. Finally, the coordinated account metrics are imperfect because we do not have ground truth labels. While different coordination metrics show the robustness of our results, these metrics are not perfect indicators of coordination. Future work is therefore needed to train models on ground truth data. It will be especially useful to detect the type of coordination (retweeting the same content, versus repeating tweets, versus sharing tweets in synchronized times, etc.) which may be a factor in how coordinated accounts behave.

Broader Perspective, Ethics and Competing Interests

All data is public and collected following Twitter's terms of service, with the study considered exempt by the authors' IRB. To minimize risk to users, all identifiable information was removed and analysis was performed on aggregated data. Data were publicly collected in the U.S. and did not require consent. We therefore believe the negative outcomes of the use of these data are minimal.

Our analysis of these data will have broad positive impact in understanding tactics of information campaigns. Researchers can use these findings to potentially better identify information campaigns in the future and reduce the harm they continue to pose. There is a chance that knowledge of

these tactics could entice bad actors to change or hide their behavior, but we believe the benefit of transparency outweighs this risk. In addition, it is possible that coordinated accounts were misclassified or that indicators were incorrect. Due to our robustness checks, and care to anonymize accounts, we believe this effect has a minimal misclassification cost. While these tweets are related to the 2017 French election, we expect our findings to generalize to other political scenarios.

Acknowledgements

Funding for this work is provided through DARPA (awards # HR0011260595 and # HR001121C0169).

References

Alhuzali, H.; and Ananiadou, S. 2021. SpanEmo: Casting Multi-label Emotion Classification as Span-prediction. In *Euro ACL*, 1573–1584.

Badawy, A.; Addawood, A.; Lerman, K.; and Ferrara, E. 2019. Characterizing the 2016 Russian IRA influence campaign. *Social Network Analysis and Mining*, 9(1): 1–11.

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.

Bail, C. A.; Guay, B.; Maloney, E.; Combs, A.; Hillygus, D. S.; Merhout, F.; Freelon, D.; and Volfovsky, A. 2020. Assessing the Russian Internet Research Agency's impact on the political attitudes and behaviors of American Twitter users in late 2017. *PNAS*, 117(1): 243–250.

Barbieri, F.; Anke, L. E.; and Camacho-Collados, J. 2022. XLM-T: Multilingual language models in twitter for sentiment analysis and beyond. In *LREC*, 258–266.

Barbieri, F.; Anke, L. E.; and Camacho-Collados, J. 2023. cardiffnlp/twitter-xlm-roberta-base-sentiment. https://huggingface.co/cardiffnlp/twitter-xlm-roberta-base-sentiment.

Baziotis, C.; Athanasiou, N.; Chronopoulou, A.; Kolovou, A.; Paraskevopoulos, G.; Ellinas, N.; Narayanan, S.; and Potamianos, A. 2018. Ntua-slp at semeval-2018 task 1: Predicting affective content in tweets with deep attentive rnns and transfer learning. *arXiv preprint arXiv:1804.06658*.

Bessi, A.; and Ferrara, E. 2016. Social bots distort the 2016 US Presidential election online discussion. *First Monday*, 21(11-7).

Bradshaw, S.; and Howard, P. N. 2019. The Global Disinformation Order: 2019 Global Inventory of Organised Social Media Manipulation.

Card, D.; Boydstun, A.; Gross, J. H.; Resnik, P.; and Smith, N. A. 2015. The media frames corpus: Annotations of frames across issues. In *ICNLP 2015*, 438–444.

Chen, C.-F.; Shi, W.; Yang, J.; and Fu, H.-H. 2021. Social bots' role in climate change discussion on Twitter: Measuring standpoints, topics, and interaction strategies. *Advances in Climate Change Research*, 12(6): 913–923.

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.

Cinelli, M.; Cresci, S.; Quattrociocchi, W.; Tesconi, M.; and Zola, P. 2022. Coordinated Inauthentic Behavior and Information Spreading on Twitter. *Decis. Support Syst.*, 160(C).

Cohen, J. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1): 37–46.

Cox, D.; and Cox, A. D. 2002. Beyond first impressions: The effects of repeated exposure on consumer liking of visually complex and simple product designs. *Journal of the Academy of Marketing Science*, 30(2): 119–130.

Duppada, V.; Jain, R.; and Hiray, S. 2018. SeerNet at SemEval-2018 Task 1: Domain Adaptation for Affect in Tweets. In *SemEval*, 18–23.

Ehrett, C.; Linvill, D. L.; Smith, H.; Warren, P. L.; Bellamy, L.; Moawad, M.; Moran, O.; and Moody, M. 2021. Inauthentic Newsfeeds and Agenda Setting in a Coordinated Inauthentic Information Operation. *Social Science Computer Review*, 0(0).

Eisenstein, J.; Ahmed, A.; and Xing, E. P. 2011. Sparse Additive Generative Models of Text. In *ICML*, 1041–1048.

Ferrara, E. 2017. Disinformation and social bot operations in the run up to the 2017 French presidential election. *First Monday*, 22(8).

Ferrara, E.; Varol, O.; Davis, C.; Menczer, F.; and Flammini, A. 2016. The rise of social bots. *CACM*, 59(7): 96–104.

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/.

Fridkin, K. L.; and Kenney, P. J. 2004. Do Negative Messages Work?: The Impact of Negativity on Citizens' Evaluations of Candidates. *American Politics Research*, 32(5): 570–605.

Garten, J.; Hoover, J.; Johnson, K. M.; Boghrati, R.; Iskiwitch, C.; and Dehghani, M. 2018. Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis. *Behavior research methods*, 50(1): 344–361.

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

Giglietto, F.; Righetti, N.; Rossi, L.; and Marino, G. 2020a. Coordinated Link Sharing Behavior as a Signal to Surface Sources of Problematic Information on Facebook. In *SMSociety*, 85–91.

Giglietto, F.; Righetti, N.; Rossi, L.; and Marino, G. 2020b. It takes a village to manipulate the media: coordinated link sharing behavior during 2018 and 2019 Italian elections. *Information, Communication & Society*, 23(6): 867–891.

Glandt, K.; Khanal, S.; Li, Y.; Caragea, D.; and Caragea, C. 2021. Stance Detection in COVID-19 Tweets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1596–1611. Online: Association for Computational Linguistics.

Graham, J.; Haidt, J.; Koleva, S.; Motyl, M.; Iyer, R.; Wojcik, S. P.; and Ditto, P. H. 2013. *Chapter Two - Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism*, volume 47, 55–130. Academic Press. ISBN 0065-2601.

Graham, J.; Haidt, J.; Motyl, M.; Meindl, P.; Iskiwitch, C.; and Mooijman, M. 2018. Moral foundations theory. *Atlas of moral psychology*, 211.

Graham, T.; Bruns, A.; Zhu, G.; and Campbell, R. 2020. *Like a virus: The coordinated spread of Coronavirus disinformation*. Canberra, A.C.T: The Australia Institute.

Gray, R. 2017. The Ḿacron LeaksŔebel in the Briefing Room. *The Atlantic*.

Guo, Y.; Rennard, V.; Xypolopoulos, C.; and Vazirgiannis, M. 2021. BERTweetFR : Domain Adaptation of Pre-Trained Language Models for French Tweets. In *W-NUT 2021*, 445–450. Online: ACL.

Hardalov, M.; Arora, A.; Nakov, P.; and Augenstein, I. 2022. A Survey on Stance Detection for Mis- and Disinformation Identification. In *NAACL 2022*, 1259–1277. Seattle, United States: Association for Computational Linguistics.

He, H.; and Xia, R. 2018. Joint Binary Neural Network for Multi-label Learning with Applications to Emotion Classification. In Zhang, M.; Ng, V.; Zhao, D.; Li, S.; and Zan, H., eds., *NLP and Chinese Computing*, 250–259. Cham.

He, Z.; Mokhberian, N.; and Lerman, K. 2022. Infusing Knowledge from Wikipedia to Enhance Stance Detection. In *WASSA*, 71–77. Dublin, Ireland: Association for Computational Linguistics.

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

Hoover, J.; and et al. 2019. Moral Foundations Twitter Corpus: A collection of 35k tweets annotated for moral sentiment.

Howard, P. N.; and Kollanyi, B. 2016. Bots,# StrongerIn, and# Brexit: computational propaganda during the UK-EU referendum. *arXiv preprint arXiv:1606.06356*.

Jelodar, H.; Wang, Y.; Yuan, C.; Feng, X.; Jiang, X.; Li, Y.; and Zhao, L. 2019. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, 78(11): 15169–15211.

Kim, Y. M. 2018. Uncover: strategies and tactics of Russian interference in US elections. *Young*, 9(04).

Kirdemir, B.; Adeliyi, O.; and Agarwal, N. 2022. Towards Characterizing Coordinated Inauthentic Behaviors on YouTube. In *ROMCIR*. Stavanger, Norway.

Küçük, D.; and Can, F. 2020. Stance detection: A survey. *CSUR*, 53(1): 1–37.

Lachat, R.; and Michel, E. 2020. Campaigning in an unprecedented election: issue competition in the French 2017 presidential election. *W Euro Politics*, 43(3): 565–586.

Li, L.; Wang, H.; Sun, X.; Chang, B.; Zhao, S.; and Sha, L. 2015a. Multi-label Text Categorization with Joint Learning Predictions-as-Features Method. In *EMNLP*, 835–839.

Li, S.; Huang, L.; Wang, R.; and Zhou, G. 2015b. Sentence-level Emotion Classification with Label and Context Dependence. In *ACL*, 1045–1053.

Li, Y.; Zhao, C.; and Caragea, C. 2021. Improving Stance Detection with Multi-Dataset Learning and Knowledge Distillation. In *EMNLP*, 6332–6345. ACL.

Linvill, D. L.; Warren, P. L.; and Moore, A. E. 2021. Talking to Trolls—How Users Respond to a Coordinated Information Operation and Why They're So Supportive. *JCMC*, 27(1). Zmab022.

Mazza, M.; Cola, G.; and Tesconi, M. 2022. Ready-to-(ab)use: From fake account trafficking to coordinated inauthentic behavior on Twitter. *OSNM*, 31: 100224.

McHugh, M. L. 2012. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)*, 22(3): 276–282.

Mei, Q.; Ling, X.; Wondra, M.; Su, H.; and Zhai, C. 2007. Topic Sentiment Mixture: Modeling Facets and Opinions in Weblogs. In *WWW*, 171–180.

Mohammad, S.; and Turney, P. 2010. Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon. In *W-CAAGET*, 26–34.

Mohammad, S. M.; Bravo-Marquez, F.; Salameh, M.; and Kiritchenko, S. 2018. SemEval-2018 Task 1: Affect in Tweets. In *SemEval*.

Morstatter, F.; Pfeffer, J.; Liu, H.; and Carley, K. 2013. Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. In *ICWSM*, volume 7, 400–408.

Nizzoli, L.; Tardelli, S.; Avvenuti, M.; Cresci, S.; and Tesconi, M. 2021. Coordinated Behavior on Social Media in 2019 UK General Election. *ICWSM*, 15(1): 443–454.

Nooralahzadeh, F.; Arunachalam, V.; and Chiru, C.-G. 2013. 2012 Presidential Elections on Twitter–An Analysis of How the US and French Election were Reflected in Tweets. In *CSCS*, 240–246.

Pacheco, D.; Hui, P.; Torres-Lugo, C.; Truong, B. T.; Flammini, A.; and Menczer, F. 2021. Uncovering Coordinated Networks on Social Media: Methods and Case Studies. *ICWSM*, 21: 455–466.

Paper, O. W. 2022. Suspicious Twitter Activity around the Russian Invasion of Ukraine.

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.

Piña-García, C. A.; and Espinoza, A. 2022. Coordinated campaigns on Twitter during the coronavirus health crisis in Mexico. *Tapuya: Latin American Science, Technology and Society*, 0(0): 2035935.

Ratkiewicz, J.; Conover, M.; Meiss, M.; Goncalves, B.; Flammini, A.; and Menczer, F. 2021. Detecting and Tracking Political Abuse in Social Media. *ICWSM*, 5(1): 297–304.

Ratkiewicz, J.; Conover, M.; Meiss, M.; Gonçalves, B.; Patil, S.; Flammini, A.; and Menczer, F. 2011. Truthy: Mapping the Spread of Astroturf in Microblog Streams. In *Companion WWW*, 249–252.

Rauchfleisch, A.; and Kaiser, J. 2020. The False positive problem of automatic bot detection in social science research. *PLOS ONE*, 15(10): 1–20.

Sayyadiharikandeh, M.; Varol, O.; Yang, K.; Flammini, A.; and Menczer, F. 2020. Detection of Novel Social Bots by Ensembles of Specialized Classifiers. In *CIKM*, 2725–2732.

Schliebs, M.; Bailey, H.; Bright, J.; and Howard, P. 2021. China's inauthentic UK Twitter diplomacy: a coordinated network amplifying PRC diplomats.

Sharma, K.; Zhang, Y.; Ferrara, E.; and Liu, Y. 2021. Identifying Coordinated Accounts on Social Media through Hidden Influence and Group Behaviours. In *KDD*, 1441–1451.

Starbird, K. 2019. Disinformation's spread: bots, trolls and all of us. *Nature*, 571(7766): 449–450.

Stokes, D. E. 1963. Spatial models of party competition. *AmE.an political science review*, 57(2): 368–377.

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

Tucker, J. A.; Guess, A.; Barberá, P.; Vaccari, C.; Siegel, A.; Sanovich, S.; Stukal, D.; and Nyhan, B. 2018. Social media, political polarization, and political disinformation: A review of the scientific literature. *Political polarization*.

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

Vilmer, J.-B. J. 2021. Fighting Information Manipulation: The French Experience. In Jayakumar, S.; Ang, B.; and Anwar, N. D., eds., *Disinformation and Fake News*, 75–89.

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

Wang, X.; Li, J.; Srivatsavaya, E.; and Rajtmajer, S. 2023. Evidence of inter-state coordination amongst state-backed information operations. *Scientific Reports*, 13(1): 7716.

Wang, Y.; and Pal, A. 2015. Detecting Emotions in Social Media: A Constrained Optimization Approach. In *IJCAI*, 996–1002.

Weber, D.; and Falzon, L. 2021. Temporal Nuances of Coordination Network Semantics.

Xu, P.; Liu, Z.; Winata, G. I.; Lin, Z.; and Fung, P. 2020. Emograph: Capturing emotion correlations using graph networks. *arXiv preprint arXiv:2008.09378*.

Ying, W.; Xiang, R.; and Lu, Q. 2019. Improving Multi-label Emotion Classification by Integrating both General and Domain-specific Knowledge. In *W-NUT*, 316–321.

Yu, J.; Marujo, L.; Jiang, J.; Karuturi, P.; and Brendel, W. 2018. Improving Multi-label Emotion Classification via Sentiment Classification with Dual Attention Transfer Network. In *EMNLP*, 1097–1102.

Ethics 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, because our work analyzes information campaigns while cautioning to avoid targeting any individual account, which may be incorrectly classified.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope?
   Yes, we avoid making claims that go beyond our statements in the introduction and abstract.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made?
   Yes, see Methods.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions?
   Yes, see Data subsection in Methods.
   (e) Did you describe the limitations of your work?
   Yes, see Discussion.
   (f) Did you discuss any potential negative societal impacts of your work?
   Yes, see Broader Perspectives, Ethics, and Competing Interests.
   (g) Did you discuss any potential misuse of your work?
   Yes, see Broader perspectives, ethics, and competing interests.
   (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 Broader perspectives, ethics, and competing interests.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them?
   Yes, we confirm the paper conforms to these guidelines.
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, we share a link to our code in the Introduction.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)?
Yes, see Methods and our code (https://github.com/KeithBurghardt/Coordination/).

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)?
Yes, see Fig. 2, 7, and 8.

(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, we state these details in appropriate subsections within the Methods section.

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

(f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance?
Yes, see Broader perspective, ethics and competing interests.

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?
NA.

(b) Did you mention the license of the assets?
NA.

(c) Did you include any new assets in the supplemental material or as a URL?
Yes, see our Github link listed in the Introduction.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating?
Yes, see Broader perspective, ethics and competing interests.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content?
Yes, see Broader perspective, ethics and competing interests.

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))?
NA

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

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?
NA.

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals?
Yes, see Broader perspective, ethics and competing interests.

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

(d) Did you discuss how data is stored, shared, and deidentified?
Yes, see Broader perspective, ethics and competing interests.