Geographical visualization of tweets, misinformation, and extremism during the USA 2020 presidential election using LSTM, NLP, and GIS

Mahdi Hashemi[1*]

*Correspondence:
mhashem2@gmu.edu

[1] Department of Information Sciences and Technology, George Mason University, 4400 University Dr, Fairfax, VA 22030, USA

Abstract

Disinformation campaigns on online social networks (OSN) in recent years, have underscored democracies' vulnerability to such operations and the importance of identifying such operations and dissecting their methods, intents, and source. With a focus on the USA 2020 presidential election, a total of 1,349,373 original Tweets have been collected by our server in real-time from the beginning of April 2020 to the end of January 2021, using four keywords: Trump, Biden, Democrats, and Republicans. In this work, deep learning, natural language processing, geographical information systems, and statistical tools are used to geographically visualize and discover if the political misinformation and extremism, political affiliation, and topics of conversations on social media are correlated with the USA 2020 presidential election results. To this end, a deep neural network is trained using 40,000 manually classified Tweets and further used to automatically classify the entire set of Tweets based on their political affiliation, topic, and whether or not they contain misinformation or extremism. It is shown that, there is a correlation between the aforementioned classes of Tweets and the election results. In other words, the political affiliation of topics and the extent of misinformation and extremism on social media are correlated with the election results to some level. The strongest correlation highlighted that the ratio of Rightist versus Leftist misinformation Tweets has a 0.67 correlation coefficient with the ratio of Trump votes versus Biden votes, across different states.

Keywords: Misinformation, Elections, Social media, Spatial visualization, Deep learning, Natural language processing

Introduction

With the business model to make profit from advertisement, online social networks (OSN) were established in early 2000s. With billions of users worldwide, sharing their personal information and views on various issues, the information value of such platforms was discovered by companies, organizations, states, and individuals. With the possibility to post any content free of fact-checking filters [1] or source verification [2] with no legal consequences, comment on other posts, no face-to-face interaction, and

stay anonymous, such platforms have become the ground for spreading misinformation and extreme opinions by individuals, organizations, and states with various goals.

OSN have become the primary discussion channel of political opinions. Their political content and the way it is spread throughout the platform has the power to alter public opinion [3, 4]. Yet, they were not invented with this aim in mind. OSN were developed to make revenue through advertisement. The higher the number of users, the time they spent on the platform, and their engagement (posting, liking, commenting, etc.), the higher the advertisement revenue. With this business model in mind, user satisfaction [5] and engagement translate to profit. In this business model, users are customers, whose satisfaction is important, while in the underlying political discourse, users are citizens [6].

OSN provide citizens with the opportunity to not only be the audience of news and opinions but to engage in a discussion, express their views by reacting and commenting [7], and participate in shaping and directing the political discourse. This new type of interactivity has posed political campaigns and other political influencers, benign or malignant, to a new set of opportunities and adversities. Because of this two-way interactivity and the possibility for citizens to enter the discussion and express their opinions, TV and newspapers are slowly giving their place as the primary source of news to OSN, websites, and blogs [8–10]. This has not been unnoticed by political campaigns who seek elections, as well as, political activists and commentators, and malign actors who attempt to influence and distort the news and its spread toward their ends and goals. Twitter, Facebook, and Instagram have observed immense presence of different political figures and political candidates who deploy OSN platforms to broadcast their activities and opinions on a wide range of national and international matters [11–14].

While misinformation refers to false information, which misleads the readers, and its unintentional spread, disinformation refers to the deliberate spread of false information (misinformation) by an entity to mislead the readers. If the false information is in line with someone's existing views, s/he is vulnerable to believing it, without questioning its source or factuality [15], because people tend to consider their perception of reality as truth [16]. Disinformation operations are conducted on OSN by individuals and states, to influence internal matters in their own country or to reach directly to citizens of another country, equip them with false information [3, 4], and thus change the direction of their political discourse, amplify their problems, or sow mistrust and anonymity among them. Disinformation operations do not target the physical infrastructure. They target the democracy's soft spot at its heart, the free speech. Democracies rely on people and disinformation feeds people with misinformation. Disinformation has slow, invisible, and complicated impacts, unlike wars which have swift, highly visible, and easily understandable impacts.

Disinformation operations were conducted by domestic actors to: impugn President Obama's religion and birthplace [17, 18], negate public opinion on Affordable Care Act [19, 20], misrepresent the evidence with regard to Iraq's role in the 9/11 attack and its mass destruction weapons [21, 22], and undermine the climate change's factuality [23, 24], and by foreign actors to cast doubt on the credibility of the U.S. political system and 2016 federal election and to polarize U.S. citizens [25, 26]. In 2018, Cambridge Analytica sent personalized political messages to U.S. citizens to influence their opinion about U.S.

internal policies [27]. Misinformation operations have been utilized by other countries as well [28–30]. Piña-García and Espinoza [31] exposed how coordinated campaigns (i.e. astroturfing) were used to influence and manipulate public opinion during the coronavirus health crisis in Mexico and provided insight into how they were detected.

All this highlights the democracy's vulnerability to disinformation operations and the importance of studies on understanding them. This study focuses on visualizing the geographical distribution of political terms, parties, misinformation, extremism, and topics among Tweets, during the USA 2020 presidential election and attempts to answer the question, whether there is any correlation between the aforementioned classes of Tweets and the election results. To this aim, 1,349,373 original Tweets have been collected in real-time from April 2020 until January 2021, based on four terms: Trump, Biden, Democrats, and Republicans. Out of these Tweets, 40,000 were manually labeled based on political affiliation, the Tweet's topic, factuality of the information within, and presence of extremism. Then a long short-term memory (LSTM) network was trained to automatically classify the entire set of Tweets into these classes. Since almost all Tweets lack the geographical tag, the location description of the user posting the Tweet was dissected using natural language processing (NLP) methods to identify which state in the USA they reside in. Geographical information systems (GIS) capabilities were deployed to tie the geographical location to the predicted information and visualize the distribution of misinformation, extremism, political affiliation, and topics across the country. Finally, correlation coefficient is calculated between the number of Tweets in the aforementioned classes in each state and the number of votes for Trump and Biden. It was shown that there is a correlation between the size of the aforementioned classes of Tweets and the election results. For instance, a higher ratio of Tweets affiliated with one party is correlated with a higher ratio of votes for that party's candidate.

The following section reviews some of the related literature to this work. Sect. "Data description" describes the data collection process. Sect. "Classification of the Textual Content of Tweets" outlines the automatic classification of Tweets based on their political affiliation, topic, and presence of misinformation and extremism. Sect. "Identifying the Geographical Location" explains how each Tweet is associated with a state in the USA. Sect. "Geographical Visualization of Labeled Tweets and Their Correlation with Election Results" provides geographical visualizations of the distribution of misinformation, extremism and topics along with their political affiliation across different states in the USA and investigates the correlation between these classes and the USA 2020 presidential election results. Sect. "Conclusions and future directions" concludes this paper and provides future research directions.

Related work

Desouza et al. [27] enumerated the following factors as to why OSN play a key role in shaping the political discourse: (a) data volume and diversity of the data sources, (b) analytical methods that extract semantic knowledge from large volumes of data,(c) automatic algorithms that learn citizens' personal views and preferences, (d) advancements in behavioral science that provides tactics for persuading humans toward particular actions [32], and (e) the ability to test and modify aforementioned techniques on OSN at a relatively low cost.

Most OSN users are passive, i.e. they read the content [33] without contributing in it [34]. Consequently, active and hyperactive users shape the OSN content. Hyperactivity is one of the information operation tactics to influence discussions by intensively contributing to OSN content [35, 36]. This could be done by both human and automated accounts. Automated accounts (aka bots which is short for robot) are autonomous software whose tasks resemble those of a human on Twitter, such as liking, tweeting, and retweeting, but for specific goals and on a large scale. They execute their actions through the Twitter API. Hyperactive users are those who over-proportionally distribute their political opinions on OSN, compared to regular users, by liking, commenting, tweeting, or other possible means. By overrepresenting the political issues and opinions that are important to them, hyperactive users deform the actual picture of public opinion on OSN and distort the public communication and discussions toward their ends. Papakyriakopoulos et al. [37], Thieltges et al. [38], and Shao et al. [39] showed that the recommendation systems in OSN not only do not prevent this distortion, but also magnify hyperactive users' interests. OSN recommendation algorithms promote most popular or most liked information. In their business model this translates to more advertisement revenue because popular content encourages users to engage more and spend more time on the platform.

Additionally, these algorithms on platforms such as Facebook, Twitter, and YouTube are designed to offer users content deemed likely to engage them. They do not offer citizens a neutral space to engage in conversations, but offer them highly contrived, personalized media experiences designed to serve the needs of advertisers [1]. This, in turn, results in feeding inflammatory content to users who are drawn to extremist narratives with continually more of the same, leading users down a rabbit hole of extremism [1, 40, 41].

Bots (fake accounts) on OSN have extensively been used to influence political campaigns by increasing a candidate's number of followers, increasing a post or hashtag's number of followers, and providing positive or negative comments on other posts [42]. As a side effect or second order effect, this also distorts the statistics reported by mainstream news media about public sentiments on different candidates and opinions.

A few methods are proposed to stop or slow down online disinformation operations through OSN and search engines. Costine [43] proposed to flag and down-rank inaccurate claims on OSN, detected by outside fact-checkers. Garrett and Weeks [44] proposed to bring contextual awareness by involving some users from the poster's social network in the fact-checking process. In other words, allowing users to correct their peers. Google is introducing two features, claim accuracy and source reliability, to be embedded in their search engine to fight disinformation operations. One focuses on informing the user that a piece of information or the retrieved result is false by having the word false mentioned next to that specific search result [45, 46] and the other focuses on down-ranking the retrieved information when its source has a low trustworthiness rate [47]. For the earlier approach, Google is encouraging outside fact-checker organizations to make their results machine-readable, so that their search engine can automatically embed the fact-checking results into search results [45, 46]. For the latter approach, Google is working on algorithms to rate the trustworthiness of the sources of search results [48].

Garrett [49] proposed to contain the spread of messages and posts that carry extreme anger, outrage, and distrust as another approach to reduce the exposure of socially harmful falsehood to mass audience, on OSN. This approach is based on the fact that disinformation operations often take advantage of emotional extremity to influence how people respond and react to their information environment. An exploratory analysis by Shu et al. [50] on multiple OSN political and entertainment datasets showed that:

- Users who post real news tend to have longer account ages than those who post false news, implying that fresh accounts are more probable than old ones to be intentionally created for spreading false news,
- Real news has more neutral replies over positive and negative replies, whereas false news has more negative replies,
- Real news tends to have a bigger ratio of likes and replies, whereas false news tends to have a bigger ratio of Retweets, and
- The number of Retweets for real news steadily increases over time, whereas that number suddenly jumps in the beginning and then remains constant over time for false news.

Some attempts have been made to automatically detect political misinformation on OSN. Among the features used in machine learning models to detect false political information are: features extracted from the textual content [2, 50–52], sentiment [53], polarity [52], subjectivity [52], disagreement [52], hashtags [2], number of replies [52], number of images, videos, question marks, exclamation points, first/second/third-person pronouns, and smile emoticons in the Tweet thread (conversation tree) [52], account age [52], user engagement features (i.e. number of replies, re-Tweets, and likes) [50, 51], features extracted from user friendship network [51, 52], and user profile features (e.g. user credibility and political affiliation) [51], topology of the diffusion network [53], and features extracted from the content of the URLs mentioned in the Tweet [2]. Among these features, the textual content, hashtags, number of images, videos, and smile emoticons in the Tweet thread (conversation tree), user engagement features, user friendship network, diffusion networks, and the content of the URLs mentioned in the Tweet have shown to be the most effective in identifying false political information.

Our work stands out because it uses both the user's location description and the Tweet's content and combines the power of deep learning, NLP, and GIS to visualize the distribution of misinformation, extremism, and topics along with their political affiliation across USA, during the USA 2020 presidential election. It also investigates any correlation between Tweets and the election results.

Data description

Twitter is chosen for this project because not only is it a prominent example of OSN [54], with currently over 330 million monthly active users [55], but also it is a public forum where everyone's posts are publicly available, and it plays an exceptional role in spreading political misinformation [27]. Original Tweets, written in English, containing at least one of the following four terms: Trump, Biden, Democrats, and Republicans have been collected by our server in real time from the beginning of April 2020 to the

Table 1 Percentage of 1,349,373 original Tweets collected from April 2020 to January 2021 containing each keyword

Trump %	Biden %	Republicans %	Democrats %
78	24	5	5

Table 2 Twitter accounts that posted the largest number of Tweets in our collection, along with their number of Tweets

@TomthunkitsMind	@TheHill	@rogue_corq	@randomtrump1
2787	1385	818	788

end of January 2021, using Python [56]. This resulted in 1,349,373 original Tweets (not Retweets or Replies). Table 1 shows the percentage of Tweets containing each of the four keywords. While 605,225 different Twitter accounts published these Tweets, 74% of these accounts posted only one Tweet, 12% posted only two Tweets, and the remaining 14% posted more than two Tweets. In other words, 14% of the Twitter accounts posted 44% of our collected Tweets. Table 2 shows the accounts posting the largest number of Tweets in our collection.

Classification of the textual content of tweets

We studied 40,000 Tweets with regard to their relevancy to the USA 2020 presidential election. The relevant Tweets were manually classified in three different ways: whether or not it contains misinformation, whether or not it contains extreme opinion, and whether it is rightist, leftist, or neutral. Misinformation Tweets are those propagating false information and news, conspiracy theories, and lies, as long as their falsehood can be determined through valid sources. Extreme opinion Tweets aim to create violence or radicalize people based on their political party, religion, race, etc., or undermine the country's political system and federal and local organizations. Leftists Tweets favor Democrats while Rightist Tweets favor Republicans [57].

Additionally, Tweets are classified based on their topic. Classes are not exclusive and a Tweet might fit into more than one topic. Two main topics among the 40,000 Tweets that were manually investigated include: Coronavirus pandemic (30.12% of Tweets) and Tweets that talk about politicians (74.95%). There are other topics, such as government policies (9.4%), USA institutions (9.58%), and elections (%10.51), but because of their small size, the machine would not be able to sufficiently get trained to automatically classify them.

Recurrent neural network (RNN) provides the possibility of classifying a text as a whole while taking both the sequence of a stream of textual information of arbitrary length and their contextual information into account. RNN captures more contextual and structural details from the text than just term frequencies, enabling it to distinguish among documents that a classifier based on term frequencies might not. The classification steps by RNN are as follow [58]:

Algorithm 1: RNN steps for text classification

1. Tokenization	The Tweet's text is broken into individual terms
2. Constructing a feature vector for each token	A feature vector is created for each token using word embeddings
3. Classifying the first token	RNN receives the feature vector for the first token and generates a class label as output. However, this is not considered as the class label for the entire Tweet yet
4. Classifying the next tokens, one by one, in the same order, until the last token in the Tweet	RNN receives the feature vector for the second token in the Tweet. It produces a class label as output. When RNN attempts to classify the second token, it has in its memory how and why it assigned a specific class label to the first token and it applies that knowledge when it attempts to classify the second token. In other words, while the first token is classified independently, the second token is classified not only based on its own feature vector, but also based on how and why the previous token was classified in a specific category Then RNN receives the feature vector for the third token in the Tweet. When it attempts to classify this feature vector, it also remembers and applies the knowledge of how it classified all the previous tokens. All tokens need to be fed to the RNN, one by one, until the last token

If a second Tweet needs to be classified, RNN first wipes its memory of how it classified the tokens from the previous Tweet. In other words, RNN classifies the first token of the second Tweet independently, regardless of the previous Tweet's classification.

Long short-term memory (LSTM) network [59] is a state-of-the-art architecture of RNN for classifying (i.e. labeling) a Tweet's textual content. An LSTM network (shown in Fig. 1) is composed of:

- An input layer (shown with x_t in Fig. 1), where the number of neurons in the input layer is equal to the number of explanatory variables in the feature space.
- One or more hidden layers (the gray area marked as memory cell in Fig. 1) which produces a hidden state (shown with h_t in Fig. 1) at every time step, and
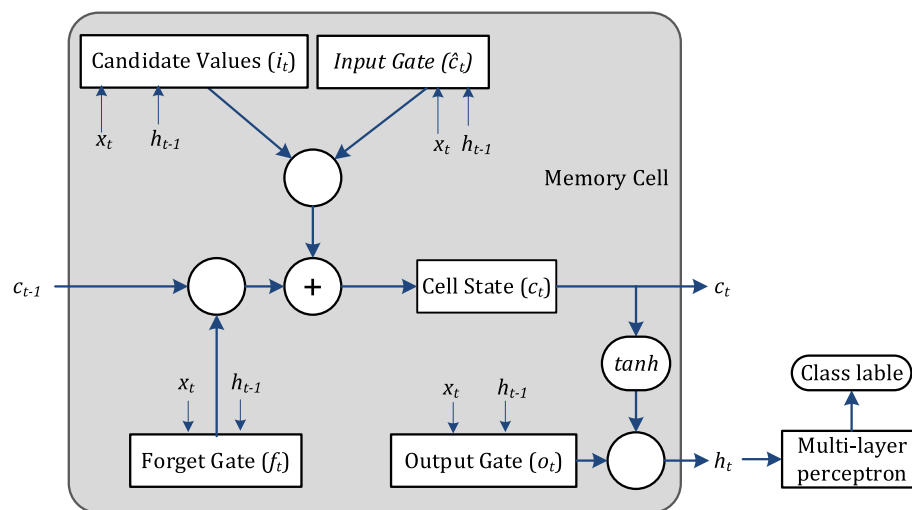

Fig. 1 Architecture of the LSTM memory cell

- A multi-layer perceptron (MLP) with softmax output layer which receives the hidden state generated by the memory cell (shown with h_t in Fig. 1) and produces a class label. There are as many units in the output layer of this MLP as there are classes. Each class is locally represented by a binary target vector with one non-zero component.

Hidden layers, also referred to as memory cells, are the main characteristic of LSTM networks. The structure of a memory cell (or hidden layer) in LSTM is illustrated in Fig. 1. Each memory cell consists of three gates whose values range from 0 to 1, acting as filters:

- The forget gate (f_t) specifies which information is erased from the cell state,
- The input gate (i_t) controls which information is added to the cell state, and
- The output gate (o_t) decides which information from the cell state is passed to the hidden state (h_t).

At every time step t, each of the three gates is presented with the input vector (x_t) at time step t as well as the output vector of the memory cells at the previous time step (h_{t-1}).

Identifying the geographical location

The Twitter server returns Tweets as JSON objects with multiple fields. The textual content of the Tweet is only one of these fields. There are four fields that are related to geographical location: geographical coordinates of the Tweet, geographical coordinates of the Twitter account owner (referred to as the user), description of the place of the Tweet, and description of the location of the account owner on their profile. The first two location fields automatically fill up if the user permits and remain empty otherwise. The other two are arbitrary texts written by the user, which can be left empty. Table 3 shows the number and percentage of Tweets that their location fields are not left empty, out of 1,349,373 original Tweets that we collected.

According to this table, the only field that can meaningfully be used to generate the location of a considerable portion of Tweets is user location description (the last column in the above table). However, writing a program that would automatically extract the state or city from this field faces multiple challenges: (a) there are duplicates or overlaps in city names, i.e. cities with the same or similar names in different states or countries, (b) it is an arbitrary text and does not follow any standards or formats, (c) cities or states are sometimes spelled in innovative or strange ways by authors, and (d) the user can write anything in this box which might not necessarily describe the user's location.

Table 3 The number and percentage of Tweets that their location fields are not left empty, out of 1,349,373 original Tweets

Tweet geographical coordinates	User geographical coordinates	Tweet place	User location
608	0	26,848	903,546
0.05%	0.00%	2.00%	66.96%

Notwithstanding these challenges, a program was written to look for certain patterns in this field and identify the state in which the user resides from the text. These patterns include the full state names, state name abbreviations, and unique city names within each state. The program was adjusted multiple times to assure that states are not mistakenly identified due to, for instance, cities having the same name in two states or countries, city names (e.g. Charlette) having an extended version that is the name of another city in another state (e.g. Charlottesville), and state abbreviations or city names that can be part of another word.

It is noteworthy that the stricter the rules, the higher the precision (correctness) of identifying the states, but the lower the recall (i.e. not identifying any state at all in many cases). By revising and adjusting the rules, we were able to identify the state for 486,969 of Tweets (recall: 53.90%), out of the 903,546 Tweets that their user location description was not empty, with a precision (or correctness) of 99%. The 99% accuracy was revealed after a manual investigation of 5000 automatically identified states.

Geographical visualization of labeled tweets and their correlation with election results

Table 4 shows the ten-fold cross-validation accuracy of the LSTM network in classifying Tweets. For political affiliation, only overall accuracy is reported since it is a three-way classification: Leftist, Neutral, and Rightist. The highest F1 score was achieved for Topic 1 (Coronavirus pandemic), followed by misinformation, Topic 2 (politicians), and extreme opinion. Next, the entire set of Tweets (1,349,373 original Tweets) are classified using the machine.
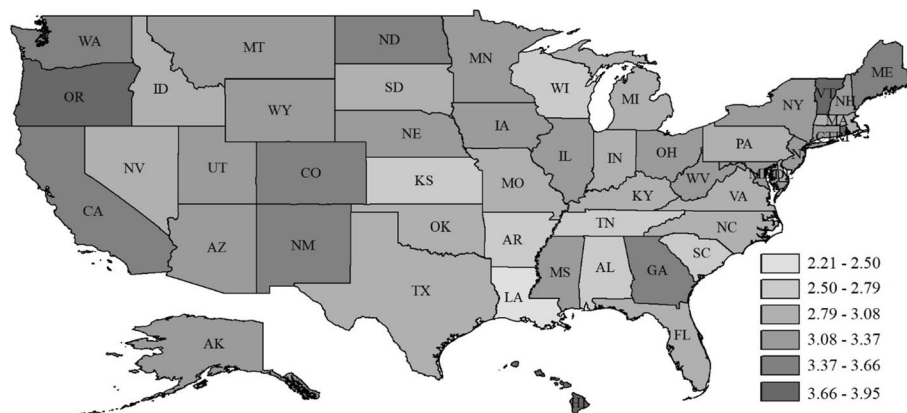
Figure 2 shows the number of times that the word 'Trump' appeared in Tweets divided by the number of times that the word 'Biden' appeared in Tweets, in each state. Figure 3 shows the collective frequency of three words: 'Coronavirus', 'Corona', or 'Covid' among Tweets divided by the collective frequency of the entire lexicon among Tweets, in each state. In other words, the percentage of the total number of words in Tweets that are either: 'Coronavirus', 'Corona', or 'Covid'.

Figure 4 shows what percentage of all Tweets were automatically labeled as misinformation in each state. Figure 5 shows the automatically detected political affiliation of misinformation Tweets in each state. Figure 6 shows what percentage of all Tweets were automatically labeled as extreme opinion in each state. Figure 7 shows the automatically detected political affiliation of extremist Tweets.

Table 4 Ten-fold cross-validation: overall accuracy, precision, recall, and F1 score in predicting misinformation, extreme opinion, political affiliation, and topics

Prediction	Overall accuracy	Precision	Recall	F1 Score
Misinformation	0.939	0.380	0.240	0.294
Extreme opinion	0.957	0.246	0.138	0.177
Political affiliation	0.570			
Topic 1 (coronavirus pandemic)	0.872	0.719	0.768	0.743
Topic 2 (politicians)	0.879	0.300	0.229	0.260

Fig. 2 The number of times that the word 'Trump' appeared in Tweets divided by the number of times that the word 'Biden' appeared in Tweets in each state, from April 2020 to January 2021
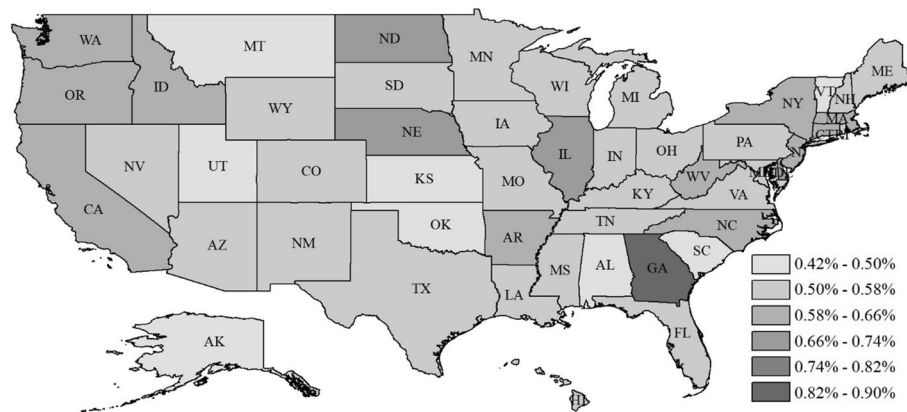

Fig. 3 Percentage of the total number of words in Tweets that are either: 'Coronavirus', 'Corona', or 'Covid' in each state, from April 2020 to January 2021
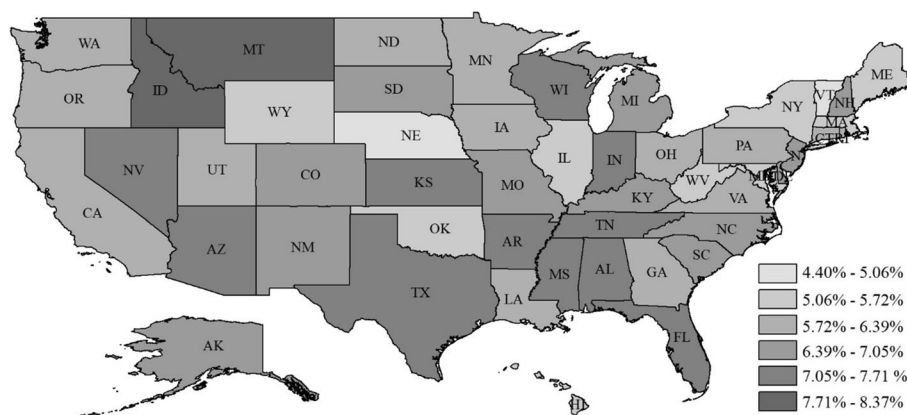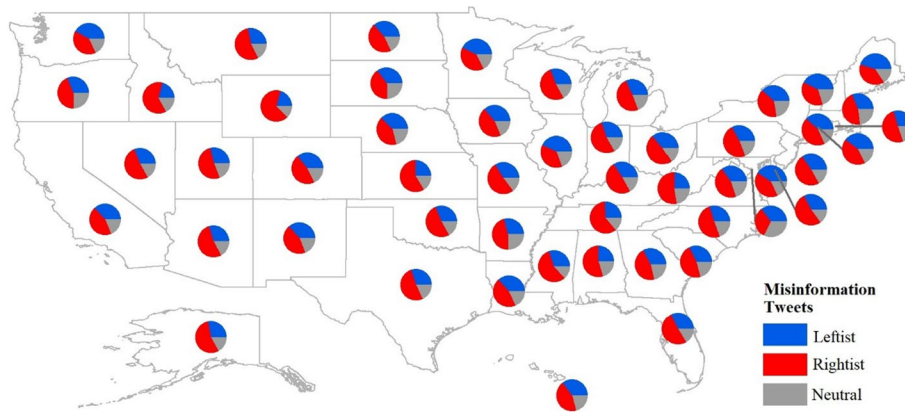

Fig. 4 Percentage of all Tweets that were automatically labeled as misinformation in each state, from April 2020 to January 2021

Figure 8 shows what percentage of all Tweets were automatically labeled as related to the Coronavirus pandemic in each state. Figure 9 shows the automatically detected political affiliation of the Coronavirus pandemic Tweets in each state. Figure 10 shows

Fig. 5 Automatically detected misinformation Tweets in each state, from April 2020 to January 2021, colored based on political affiliation
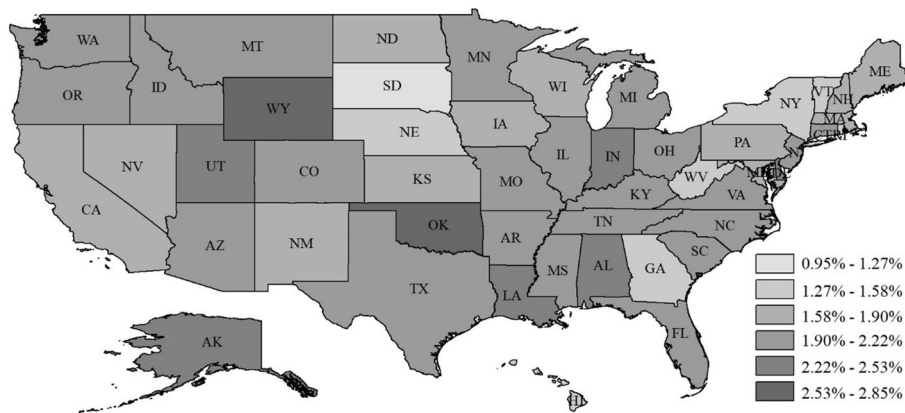

Fig. 6 Percentage of all Tweets that were automatically labeled as extreme opinion in each state, from April 2020 to January 2021
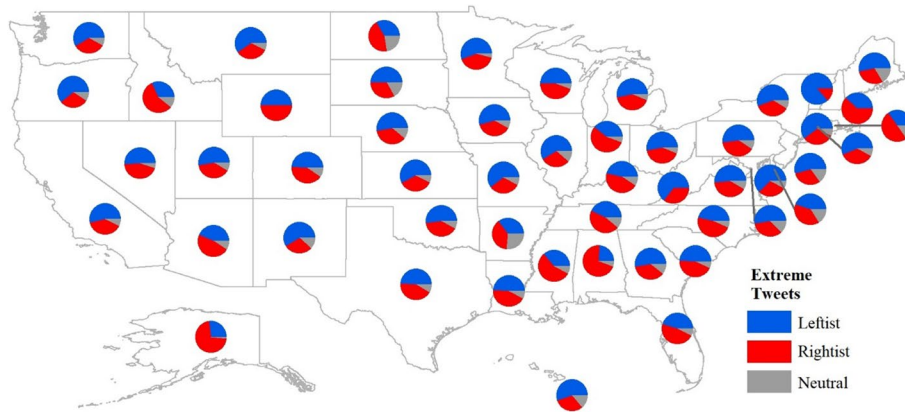

Fig. 7 Automatically detected extreme opinion Tweets in each state, from April 2020 to January 2021, colored based on political affiliation
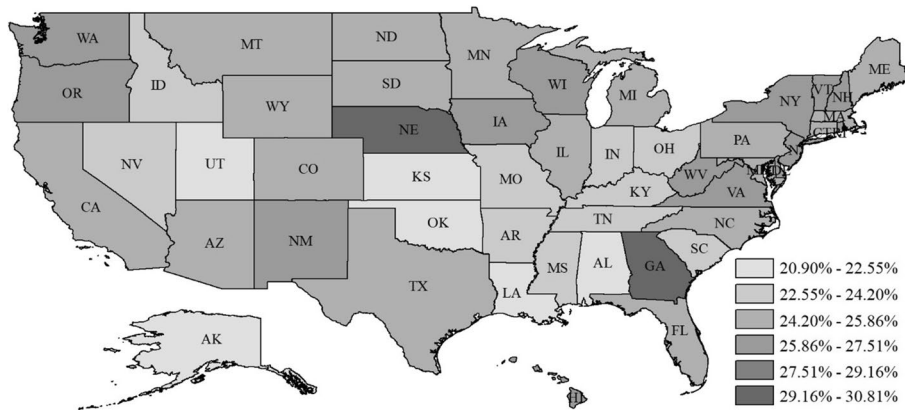
Fig. 8 Percentage of all Tweets that were automatically labeled as related to the Coronavirus pandemic in each state, from April 2020 to January 2021
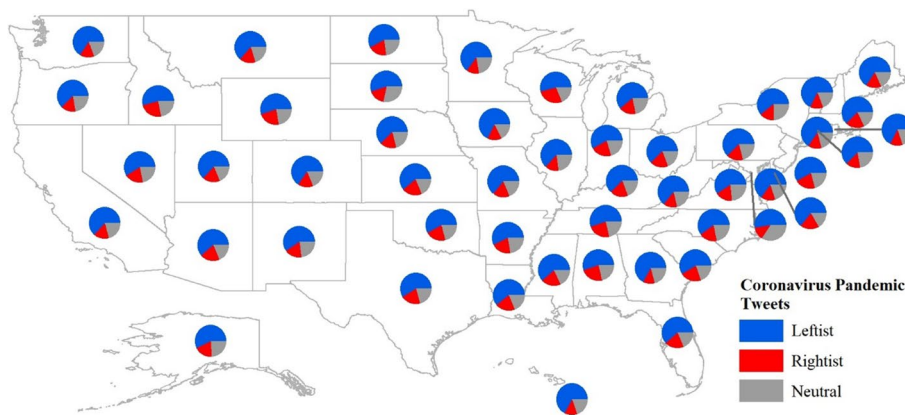

Fig. 9 Automatically detected Tweets about the Coronavirus pandemic in each state, from April 2020 to January 2021, colored based on political affiliation
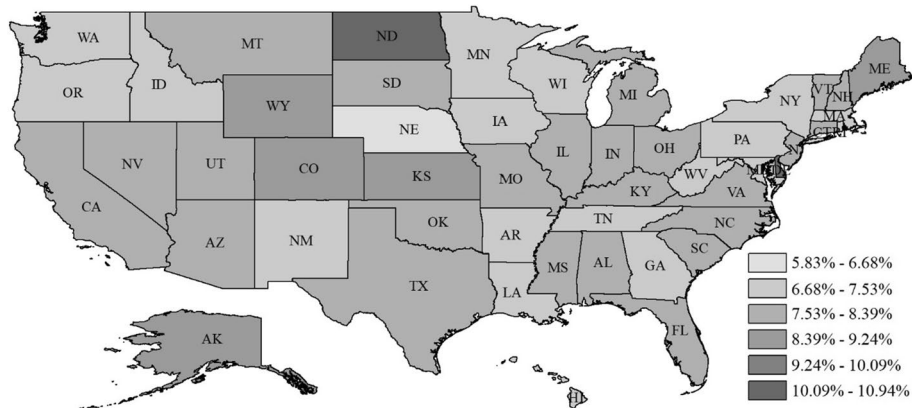

Fig. 10 Percentage of all Tweets that were automatically labeled as related to politicians in each state, from April 2020 to January 2021
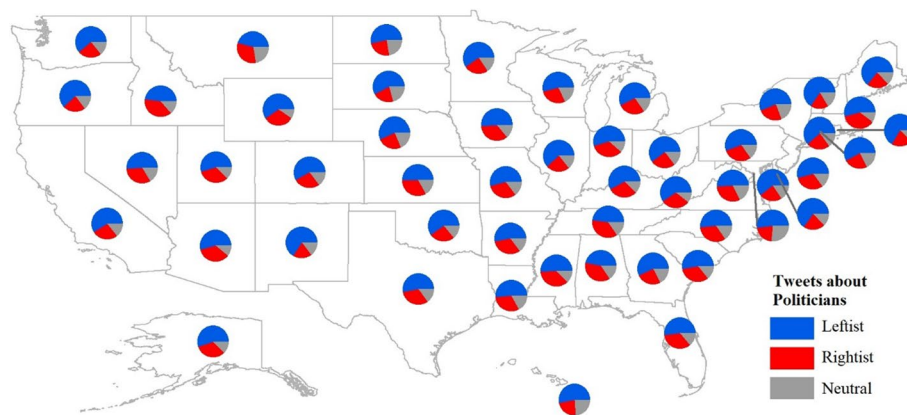
Fig. 11 Automatically detected Tweets about politicians in each state, from April 2020 to January 2021, colored based on political affiliation

Table 5 Correlation coefficient between Tweet analytics (rows) and USA 2020 presidential election results (columns); in this correlation matrix darker brown indicates higher positive correlation and darker blue indicates higher negative correlation

	Trump Votes	Biden Votes	Trump votes divided by Biden Votes
Leftist Misinformation Tweets	-0.32	0.33	-0.51
Rightist Misinformation Tweets	0.54	-0.53	0.45
Rightist Extremism Tweets	0.37	-0.37	0.33
Leftist Tweets about Coronavirus pandemic	-0.48	0.48	-0.50
Rightist Tweets about Coronavirus pandemic	0.43	-0.42	0.39
Neutral Tweets about Coronavirus pandemic	-0.45	0.45	-0.25
Rightist Tweets about Politicians	0.61	-0.6	0.52
Misinformation Tweets	0.35	-0.34	0.21
Extremism Tweets	0.44	-0.44	0.38
Tweets about Coronavirus pandemic	-0.48	0.49	-0.42
Tweets about Politicians	0.28	-0.29	0.26
Covid term frequency percentage among Tweets	-0.24	0.26	-0.21
Trump term frequency divided by Biden term frequency among Tweets	-0.31	0.29	-0.30
Rightist divided by Leftist Misinformation Tweets	0.62	-0.63	0.67
Rightist divided by Leftist Extremism Tweets	0.36	-0.36	0.33
Rightist divided by Leftist Tweets about Coronavirus pandemic	0.54	-0.53	0.53
Rightist divided by Leftist Tweets about politicians	0.51	-0.50	0.43

what percentage of all Tweets were automatically labeled as related to politicians in each state. Figure 11 shows the automatically detected political affiliation of Tweets about politicians in each state.

We obtained the number of votes for Trump and Biden in each state in the USA 2020 presidential election via [60]. Table 5 shows the correlation coefficient between the relative number of automatically classified Tweets in each category (the class size divided by the total number of Tweets in each state) and the relative number of votes for each candidate (the number of votes for each candidate divided by the total number of votes in each state). Correlation coefficients between -0.2 and 0.2 are not shown. The most significant observation is the impact of misinformation on the election results. While Leftist misinformation has a positive correlation with Biden votes and a negative correlation with Trump votes, Rightist misinformation has a positive correlation with Trump votes and a negative correlation with Biden votes. The ratio of Rightist versus Leftist misinformation Tweets has a 0.67 correlation coefficient with

the ratio of Trump votes versus Biden votes. This highlights that political misinformation on social media is fairly correlated with who people ultimately vote for. A similar observation is made with regard to extremist Tweets, but with a much lower correlation coefficient. The ratio of Rightist versus Leftist extremist Tweets has a 0.33 correlation coefficient with the ratio of Trump votes versus Biden votes.

The correlation between Tweet topics and votes is also noteworthy. While the topic of Coronavirus had a positive correlation with Biden votes, the topic of politicians had a positive correlation with Trump votes. The number of Tweets about the Coronavirus pandemic has a -0.42 correlation coefficient with the ratio of Trump votes versus Biden votes. The number of Tweets about politicians has a 0.26 correlation coefficient with the ratio of Trump votes versus Biden votes. Considering the political affiliation of Tweets will provide a clearer picture of how they are correlated with the election results. Rightist Tweets on either topic have a positive correlation with Trump votes and Leftist Tweets on either topic have a positive correlation with Biden votes. The ratio of Rightist versus Leftist Tweets about the Coronavirus pandemic has a 0.53 correlation coefficient with the ratio of Trump votes versus Biden votes. The ratio of Rightist versus Leftist Tweets about politicians has a 0.43 correlation coefficient with the ratio of Trump votes versus Biden votes.

Conclusions and future directions

In this paper, machine learning, natural language processing, geographical visualization, and statistical tools were used to understand the relationship between political misinformation and extremism, topics, and their political affiliation among conversations on social media on one hand, and the USA 2020 presidential election results on the other hand. After automatic classification of Tweets based on the aforementioned classes, it was shown that there is a correlation between these factors and the election results. The strongest correlation highlighted that the ratio of Rightist versus Leftist misinformation Tweets has a 0.67 correlation coefficient with the ratio of Trump votes versus Biden votes. A similar result but with a correlation coefficient of 0.33 was obtained with regard to extremism. Rightist Tweets about the Coronavirus pandemic or politicians were found to have a positive correlation with Trump votes and Leftist Tweets on either topic had a positive correlation with Biden votes. The prevalence of Tweets about the Coronavirus pandemic had a positive correlation with Biden votes and the prevalence of Tweets about Politicians had a positive correlation with Trump votes. This indicates that there is a correlation between what happens on Twitter and how people vote, however, this is not a causal inference. In other words, it is not known whether it is topics, misinformation, and extremism on social media that drive how people vote or vice versa. Was the misinformation and extremism on social media that convinced people to vote one way or the other, or people had already made their mind before engaging in online conversations? Did the topic of Tweets on social media caused people to vote one way or the other, or was it people's already decided votes that made them engage in certain conversations on social media? While these questions are deferred to another study, we showed that the political affiliation of topics and the extent of misinformation and extremism on social media are correlated with the election results to some level.

Declarations

Ethics approval and consent to participate
Not Applicable.

Consent for publication
The authors give their consent for publication.

Competing interests
The author declares no competing interests.

References
1. Ehrenfeld D, Barton M. Online public spheres in the era of fake news: implications for the composition classroom. Comput Compos. 2019;54: 102525.
2. V. Qazvinian, E. Rosengren, D. R. Radev and Q. Mei, Rumor has it identifying misinformation in microblogs in In Proceedings of the Conference on empirical methods in natural language Processing, Edinburgh, Scotland, UK, 2011.
3. S. Krishnan, J. Patel, M. J. Franklin and K. Goldberg, A methodology for learning, analyzing, and mitigating social influence bias in recommender systems in In Proceedings of the 8th ACM Conference on recommender systems, 2014.
4. D. Cosley, S. K. Lam, I. Albert, J. A. Konstan and J. Riedl Is seeing believing? how recommender system interfaces affect users' opinions in In Proceedings of the SIGCHI Conference on human factors in computing systems 2003.
5. N. Shi, M. K. Lee, C. M. Cheung and H. Chen, "The continuance of online social networks: how to keep people using Facebook? in In Proceedings of the 43rd Hawaii International Conference on system sciences 2010.
6. Sunstein CR. Republic divided democracy in the age of social media. New Jersey: Princeton University Press; 2018.
7. McCombs ME, Shaw DL. The agenda-setting function of mass media. Public Opin Q. 1972;36(2):176–87.
8. J. Gottfried and E. Shearer. Americans' online news use is closing in on TV news use 2017.
9. Shu K, Sliva A, Wang S, Tang J, Liu H. Fake news detection on social media: a data mining perspective. ACM SIGKDD Explor Newsl. 2017;19(1):22–36.
10. Hashemi M. Discovering social media topics and patterns in the Coronavirus and election era. J Inform Commun Ethics Soc. 2021. https://doi.org/10.1108/JICES-04-2021-0039.
11. Hegelich S, Shahrezaye M. The communication behavior of German MPs on Twitter: preaching to the converted and attacking opponents. Eur Policy Anal. 2015;1(2):155–74.
12. Enli GS, Skogerbø E. Personalized campaigns in party-centred politics: Twitter and Facebook as arenas for political communication. Inf Commun Soc. 2013;16(5):757–74.
13. Arnaboldi V, Passarella A, Conti M, Dunbar R. Structure of ego-alter relationships of politicians in Twitter. J Comput-Mediat Commun. 2017;22(5):231–47.
14. Serrano JCM, Hegelich S, Shahrezaye M, Papakyriakopoulos O, Social Media Report. The 2017 German federal elections. Munich: TUM University Press; 2018.
15. Nickerson RS. Confirmation bias: a ubiquitous phenomenon in many guises. Rev Gen Psychol. 1998;2(2):175–220.
16. A. Ward, L. Ross, E. Reed, E. Turiel and T. Brown. Naive realism in everyday life: Implications for social conflict and misunderstanding. Values and Knowledge, pp. 103–135, 1997.
17. Hartman TK, Newmark AJ. Motivated reasoning, political sophistication, and associations between president Obama and Islam. PS Polit Sci Polit. 2012;45(3):449–55.
18. YouGov Staff, 15% 'know for sure' Obama was not born in the U.S. Economist/YouGov poll, YouGov, 12 February 2014. Available: https://today.yougov.com/topics/politics/articles-reports/2014/02/12/know-for-sure.
19. Meirick PC. Motivated misperception? Party, education, partisan news, and belief in "death panels." J Mass Commun Quarterly. 2013;90(1):39–57.

20. Nyhan B. Why the death panel myth wouldn't die: misinformation in the health care reform debate. Forum. 2010;8(1):5.
21. Prasad M, Perrin AJ, Bezila K, Hoffman SG, Kindleberger K, Manturuk K, Powers AS. There must be a reason": Osama, Saddam, and inferred justification. Sociol Inq. 2009;79(2):142–62.
22. World Public Opinion. Percentage of Americans believing Iraq had WMD rises 9 August 2006. Available: http://world publicopinion.net/percentage-of-americans-believing-iraq-had-wmd-rises/.
23. S. Jerving, K. Jennings, M. Hirsch and S. Rust What Exxon knew about the earth's melting Arctic, Los Angeles Times, 2015.
24. Oreskes N, Conway EM. Merchants of doubt: How a handful of scientists obscured the truth on issues from tobacco smoke to global warming. New York: Bloomsbury Publishing; 2010.
25. Inkster N. Information warfare and the US presidential election. Survival. 2016;58(5):23–32.
26. Office of the Director of National Intelligence AssessingRussian activities and intentions in recent US ElectionsIntelligence Community Assessment. 2017.
27. Desouza KC, Ahmad A, Naseer H, Sharma M. Weaponizing information systems for political disruption: the actor, lever, effects, and response taxonomy (ALERT). Comput Secur. 2020;88: 101606.
28. Andre V. The Janus face of new media propaganda: The case of Patani Neojihadist YouTube warfare and its Islamophobic effect on cyber-actors. Islam Christian-Muslim Relations. 2014;25(3):335–56.
29. Molony T. Social media warfare and Kenya's conflict with Al Shabaab in Somalia: A right to know? Afr Aff. 2018;118(471):328–51.
30. Hashemi M, Hall M. Detecting and classifying online dark visual propaganda. Image Vis Comput. 2019;89(1):95–105.
31. Piña-García CA, Espinoza A. "Coordinated campaigns on Twitter during the coronavirus health crisis in Mexico Tapuya: Latin American science. Technol Soc. 2022;5(1):2035935.
32. Cole K. Turning cyberpower into idea power: the role of social media in us strategic communications school of advanced air and space studies. Alabama: Air University; 2011.
33. F. Benevenuto, T. Rodrigues, M. Cha and V. Almeida characterizing user behavior in online social networks in In Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement 2009.
34. D. M. Romero, W. Galuba, S. Asur and B. A. Huberman. Influence and passivity in social media in In Joint European Conference on machine learning and knowledge discovery in databases 2011.
35. S. Hegelich and D. Janetzko, Are social bots on Twitter political actors? Empirical evidence from a Ukrainian social botnet in In Proceedings of the Tenth International AAAI Conference on Web and Social Media. 2016.
36. J. Weedon, W. Nuland and A. Stamos information operations and Facebook," Facebook 2017.
37. Papakyriakopoulos O, Serrano JCM, Hegelich S. Political communication on social media: a tale of hyperactive users and bias in recommender systems. Online Soc Networks Media. 2020;15: 100058.
38. Thieltges A, Papakyriakopoulos O, Serrano JCM, Hegelich S. Effects of social bots in the iran-debate on Twitter. arXiv. 2018;1805:10105.
39. Shao C, Ciampaglia GL, Varol O, Yang K-C, Flammini A, Menczer F. The spread of low-credibility content by social bots. Nat Commun. 2018;9(1):4787.
40. M. Fisher and A. Taub. In search of Facebook's heroes, finding only victims the new york times. 2018.
41. Z. Tufekci. YouTube, the great radicalizer. The New York Times 2018.
42. J. Echeverria and S. Zhou. Discovery, retrieval and analysis of the 'star wars' botnet in Twitter in In Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining, 2017.
43. J. Constine. Facebook now flags and down-ranks fakenews with help from outside fact checkers 15 December 2016. Available: https://techcrunch.com/2016/12/15/facebook-now-flags-and-down-ranks-fake-news-with-help-from-outside-fact-checkers/.
44. R. K. Garrett and B. E. Weeks. The promise and peril of real-time corrections to political misperceptions in In Proceedings of the 2013 Conference on computer supported cooperative work. 2013.
45. J. Kosslyn and C. Yu, Fact check now available inGoogle Search and news around the world 7 April 2017. Available: https://www.blog.google/products/search/fact-check-now-available-google-search-and-news-around-world/.
46. E. Weise. We tried Google's new fact-check filter onthe Internet's favorite hoaxes USA Today. 2017.
47. Hodson H. Nothing but the truth. New Sci. 2015;225(3010):24.
48. Dong XL, Gabrilovich E, Murphy K, Dang V, Horn W, Lugaresi C, Sun S, Zhang W. Knowledge-based trust esti-mating the trustworthiness of web sources. ArXiv. 2015;1502:03519.
49. Garrett RK. The "echo chamber" distraction: disinformation campaigns arethe problem, not audience fragmentation. J Appl Res Mem Cogn. 2017;6:370–6.
50. Shu K, Mahudeswaran D, Wang S, Lee D, Liu H. Fake news net a data repository with news content social context and spatiotemporal information for studying fake news on social media. ArXiv. 2019;1809:012862.
51. K. Shu, S. Wang and H. Liu Beyond News Contents The Role of Social Context for Fake News Detection in In Proceedings of the 12th ACM International Conference on Web Search and Data Mining, Melbourne, VIC, Australia, 2019b.
52. C. Buntain and J. Golbeck. Automatically identifying fake news in popular twitter threads in IEEE International Conference on Smart Cloud. 2017.
53. Ratkiewicz J, Conover M, Meiss M, Gonçalves B, Patil S, Flammini A, Menczer F. Detecting and tracking the spread of astroturf memes in microblog streams ArXiv preprint. arXiv. 2010;1011:3768.
54. Chou W-YS, Hunt YM, Beckjord EB, Moser RP, Hesse BW. Social media use in the United States: implications for health communication. J Med Internet Res. 2009;11(4): e48.
55. Twitter, 2019. Available: https://about.twitter.com/company. Accessed 01 February 2019.
56. Available. https://drive.google.com/file/d/1sbqLVcDFPuG6LZRBlv6WqjgeFl1NLRn7/view?usp=sharing.
57. Hashemi M. A data-driven framework for coding the intent and extent of political tweeting, disinformation, and extremism. Information. 2021;12(4):148.
58. Hashemi M. Web page classification: a survey of perspectives, gaps, and future directions. Multimedia Tools Appl. 2020. https://doi.org/10.1007/s11042-019-08373-8.
59. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735–80.
60. Politico 2022. https://www.politico.com/2020-election/results/president. [Accessed 2022].

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.