polish sociological review 1(229)'25
ISSN 1231 – 1413
DOI:10.26412/psr229.04

PAWEŁ MATUSZEWSKI
Collegium Civitas

MICHAŁ RAMS-ŁUGOWSKI
University of Silesia in Katowice

In Search of a Conspiracy

A Practical Guide for Identifying Conspiracy Theories in Unstructured Textual Data

Abstract: Our study aims to describe a computer-aided method of searching for conspiracy thinking in unstructured textual data. Collecting such data from the Internet usually involves using keywords to find relevant documents for further analysis. Although this step determines the results, many researchers select keywords arbitrarily without evaluating their tools. We introduced a method of keyword expansion that combines word embeddings and human cognitive abilities to identify potential keywords. In our study, we found that the relatively informed participants (N = 154) could not recall even a short list of relevant keywords, and the ones they selected were mostly useless in detecting conspiracy thinking. The designed Conspiracy Thinking Index performed better in detecting conspiracy-related text in a large text corpus (≈ 1.1M tweets) than supervised machine learning algorithms while remaining simple and transparent.

Keywords: conspiracy theories, text analysis, word embeddings, keyword searching, conspiracy thinking detection

Introduction

Collecting textual data from the internet typically involves employing keywords to identify relevant documents. Most web-based tools require such keywords to narrow down a vast number of documents to a specific dataset. However, arbitrarily selecting keywords can have serious implications for research results. Different keywords can yield vastly different datasets, which can ultimately impact the producibility of research results. This article aims to address these two issues.

First, the challenge of recalling numerous words connected to a specific concept has proven difficult for humans, as demonstrated by studies conducted by Furnas et al. (1987) and Hayes & Weinstein (1990). King et al.'s (2017) recent research highlights that researchers should not select keywords without following a proper scientific procedure, as their choices tend to be arbitrary and lead to various datasets. We conducted a survey to verify this claim in the context of conspiracy theories, hypothesizing that the results for highly complex, multitopic narratives such as pandemic conspiracy theories would be even more biased compared to fairly distinctive phenomena like Obamacare. Additionally, we used our survey results as an example of the wisdom of crowds (Surowiecki 2005)

to determine if collective keyword guessing would yield better results than the method introduced in this paper.

Our second concern stems from our initial issue. We observed many academic papers that employed arbitrary and unverified keywords, which raises questions about the adequacy of classification. It is crucial to determine if the keyword list is complete, if they accurately define the subject of study, if all relevant data is collected, and if any documents are irrelevant. While researchers' intuition is valuable, it is inherently subjective and cannot be relied upon as an objective measure. In this paper, we propose a semi-automated pipeline to identify concepts, such as conspiracy, in large unstructured textual datasets. Our method initially depends on researchers' intuition, but then utilizes automated text analysis techniques and data-driven procedures to support researchers' decisions.

Based on our literature review, we perceive the arbitrariness of keyword selection as parallel to the relative rarity of studies where conspiracy theories were explicitly and clearly defined (Mahl et al. 2023). In both cases, researchers try to gather as many examples of the given phenomenon as possible instead of conceptualizing the object of their studies as such. Such attempts have already been criticized by Hegel, who called them "an aggregate of information, which has no right to bear the name of Science" in *Phenomenology of Spirit* (1977: 1).

We selected conspiracy theories as an illustrative example of an extreme case in keyword and document set discovery (Seawright & Gerring 2008). This choice was based on the fact that conspiracy theories are challenging to define, as various criteria can be employed to identify them within large text corpora (DiFonzo 2019; Uscinski 2020; Walker 2019). Thus, our method started with a clear operational definition. Additionally, conspiracy theories are a multifaceted research subject, as they are not only related to a specific topic (e.g., 5G vaccines), but also to how people discuss them. That is, while conspiracy theories can be grouped according to their subject (e.g., flat-Earth conspiracy, Kennedy conspiracy, space alien conspiracy), simply identifying a topic or attitude related to conspiracy theories may not be enough to detect conspiracy thinking. For instance, claiming that COVID-19 vaccines are dangerous could be the result of misinformation, but not necessarily a conspiracy theory. On the other hand, claiming that COVID-19 vaccines were intentionally designed by a group of satanists in power to kill as many Slavic people as possible is a conspiracy theory according to our operational definition. While vaccines are the primary focus of both examples, they do not necessarily serve as a distinguishing factor between texts that contain conspiracy theories and those that do not. Identifying conspiracy theories in large textual data appears to be a more complex task than detecting texts related to economics, for example. Additionally, some conspiracy theories have gained widespread recognition and are often used to mock the underlying ideas, making it challenging to differentiate between genuine beliefs and ridicule. Conspiracy theories can be likened to fan fiction, with new plots, events, and actors constantly emerging, and believers intentionally altering their language to evade detection by social media algorithms (Guzek 2021). Therefore, an effective method for detecting conspiracy theories should account for their diversity and ability to evolve over time.

This paper offers a comprehensive guide to selecting evidence-based keywords that extract complex and evolving concepts from large text corpora. Although several advanced statistical techniques are used, we aim to make the procedure transparent and understand-

able for readers without a solid statistical background. Our method focuses on selecting relevant documents where textual data is the primary source of information, making it particularly useful for researchers collecting data from various online sources, including those that do not allow the identification of the spreaders of information. To show usefulness of our approach, we compare it with keywords selected based on survey results and machine learning models. These points are translated into the following research questions:

RQ1. Is the introduced method a more accurate procedure of finding conspiracy theories than keyword search based on wisdom of crowds?

RQ2. Is the introduced method a more accurate procedure of finding conspiracy theories than supervised models?

Literature Review

Defining conspiracy theories

Identifying conspiracy theories in any form of communication necessitates a precise definition of the phenomenon. For our research, this requires codifying criteria that can be operationalized. An improper choice at this stage may lead to divergent results from our goal. We based our criteria on an in-depth review of existing definitions and discussions surrounding the construction of the term (Łukowski 2016; Mahl et al. 2023; Napolitano & Reuter 2021; Uscinski 2020; Walker 2018).

The term 'conspiracy theory' was properly introduced into social science by Karl Popper ([1945]2013) and later developed by Richard Hofstadter ([1965]1996). While the former described it as a phenomenon of cognitively faulty formulation of judgments about social phenomena, the latter gave it a more psychological dimension, considering conspiracy theories to be a manifestation of a 'paranoid style of thinking'. The term itself is typically viewed as derogatory and dismissive in both English (Napolitano & Reuter 2021) and Polish (Łukowski 2016) (the language of our research subject), which introduces a semantic bias and epistemic contamination in research. Focusing on the content itself produces an aggregation of thematically connected statements, which barely moves us towards an understanding of processes and structures behind the phenomenon of conspiracy theories. Therefore, we decided to use an auxiliary typology of definitions.

As Łukowski (2016) argues, firstly we must distinguish between intuitive-content-based and non-content-based (structural) approaches. The first one—which we observe in definitions proposed by Popper (2013), Zonis & Josep (1994) and Vermeule & Sunstein (2009)—focuses on collecting concrete examples, the set of which has been already based upon an arbitrary understanding of what a conspiracy theory is. Such an approach is similar to what Napolitano & Reuter (2021: 3) call a 'descriptive conceptual analysis.'

On the other hand, Łukowski (2016) proposes a way of defining conspiracy theories based on the processes behind them. He encourages researchers to study the origins of conspiracy thinking. Likewise, Napolitano & Reuter (2021: 3) propose 'conceptual engineering'—the approach of improving "on the ordinary concept by defining conspiracy theory in a way that serves a certain theoretical or practical goal."

Following the idea of conceptual engineering as well as Łukowski's (2016) remarks on the semantics of the term 'conspiracy theory', we have decided to follow an approach inspired by Lewandowski and Cook (2020) and search for 'conspiracy thinking'—more specifically, for its manifestations in textual communication—that is: to try to identify whether such an epistemic approach stands behind a given text. Ultimately, our method is intended to investigate the propagation and dynamics of conspiracy theories as such, and not as a theme or narrative independent of their actual followers. We require our algorithm to identify communications that manifest one's involvement in a given conspiracy theory. By taking up the notion of 'conspiracy thinking', we do not aim to investigate the thought process itself or arrive at epistemological, cognitive, or psychological claims. It is a methodological ploy that allows us to exclude false positives, such as criticism or ridicule of conspiracy theories. This is a similar problem to the one that has forced Czech's (2019) study to distinguish between two attitudes towards conspiracy narratives: supporting and critical. Similarly, our algorithm is designed to identify statements that are most likely to have a genuine commitment to the conspiracy theories behind them.

Using the aforementioned literature review and especially the available meta-analyses of conspiracy theories studies (Goreis & Voracek 2019; Mahl et al. 2023; Pilch et al. 2023; Stasielowicz 2022), we have developed a definition of conspiracy thinking through a content- and structure-oriented approach. This definition describes conspiracy thinking as an interpretation of history and singular events characterized by exaggerated mistrust, resistance to evidence, and possible self-contradiction. The central theme of conspiracy thinking is the opposition between 'Us' and 'Them,' where 'Us' represents perspicacious victims of the natural and righteous order and 'Them' represents clandestine and powerful villains driven by corruption and malice. Conspiracy thinking employs strategies of self-sealing by referring to unverifiable proof or undermining its critique as being part of the conspiracy.

We identified criteria for conspiracy thinking in unstructured textual data by operationalizing our definition, which included the following: (1) statements about belief in a conspiracy involving people in power or a secret group controlling the economy/politics/society, such as 5G, chemtrails, flat Earth, climate denialism, politicians being paid by foreign governments, COVID not existing, etc.; (2) statements that imply the author shares beliefs produced by existing conspiracy thinking; (3) events being explained by conspiracies and wicked intentions, such as doctors being paid off to hide the truth about a fake pandemic for profits; (4) questioning mainstream interpretations and providing alternative conspiracy explanations with a low probability; (5) stating strong beliefs about important events that are contradictory, incoherent, or unverifiable; (6) taking on the role of a victim of the mainstream narrative explaining important events; (7) statements about important events pointing to evidence resistance and self-sealing, such as believing that if NASA denies something, it's proof of a conspiracy; and (8) statements containing an extreme degree of suspicion, preventing belief in anything that doesn't fit the conspiracy theory. The criteria can be divided into two groups. 1–4 address content suggesting that the author shares conspiracy beliefs. These criteria often apply individually, particularly in shorter texts like social media posts. 5–8 are auxiliary criteria and focus on the form of argumentation, serving as cues in ambiguous cases.

Approaches to identifying conspiracy theories in unstructured textual data

There are several methods for identifying conspiracy theories (CT) on the internet, including linguistic, psychological, rhetorical, and media studies. In our literature review and analysis of two systematic literature reviews (Mahl et al. 2023; Marcellino et al. 2021), we identified five research approaches for identifying CT online based on indicators used by researchers. These include looking for 1) keywords, 2) narratives, 3) topics, 4) rhetorical strategies, and 5) online behavior of social media accounts that spread CT. Some studies use a hybrid approach that combines these perspectives.

The first method we explored is linguistic, where keywords are considered as the fundamental units of research (Houli et al. 2021; Tyagi & Carley 2021). This approach assumes that texts containing specific keywords may be classified as CT and can be manually checked. Our research follows this method by identifying CT through sets of keywords that are statistically more prevalent in CT texts. This approach can be easily automated to scan entire databases for keywords or analyze more intricate relationships between words (Grimmer & Stewart 2013).

The next strategy involves searching for CT narratives (Samory & Mitra 2018; Shahsavari et al. 2020; Tangherlini et al. 2020). These narratives are characterized by the relationships between actants, actions, and attributes or targets. By identifying these networks of words, the narrative approach helps to find the most likely representation of CT. However, this approach may require extensive and structured text bodies, such as entire internet forums, to function effectively.

The third approach is based on a search for topics, i.e. identifying texts as CTs due to the subject matter discussed. Here topics can be identified in two ways. The first one is manual, where texts are coded according to fixed criteria of CT (Kou et al. 2017) or undergo discourse analysis (Poberezhskaya 2018). The other, an automated one, uses topic models and assumes that statistical properties of textual data indicate conspiracy thinking within texts (Kant et al. 2022).

Another approach focuses on the rhetorical tactics employed by CTs and analyzes the argumentation used by their supporters (Glowacki & Taylor 2020; Nugier et al. 2018).This approach is more labor-intensive than other techniques because it involves qualitative text analysis, which is difficult to automate. However, it also requires keywords to identify, filter, and collect texts for analysis (Nugier et al. 2018).

In the last approach, the investigations concern CT-disseminating social media accounts or groups of accounts and their mutual interactions. These studies use methods such as observation, netnography or Social Network Analysis to construct a model of CT supporters' profiles and their characteristic online behavior (Ahmed et al. 2020; Bessi et al. 2015; Shahsavari et al. 2020). Although they focus on behavioral patterns of users' communication, the identification of users requires additional text analyses. Also, in this case, delivering a list of keywords specific to the CT language can prove very useful here, since not everything that is communicated by such accounts is conspiracy related.

Among the studies representing the above approaches, we encountered a number of issues that our method addresses. The first is the arbitrary selection of keywords, in which researchers use undefined criteria or where recall and precision or other evaluation metrics

resulting from such a selection are not even considered (Havey 2020; Shahsavari et al. 2020). As a result, the collected data may consist of a large proportion of false positive cases or omit cases that are relevant. Second, while some studies equate keywords with hashtags (Ahmed et al. 2020; Giachanou et al. 2021; Kant et al. 2022), this approach has limitations as not all social media users use hashtags. Hashtags may be used to artificially expand content reach or misleadingly present it as CT-related. Finally, some social platforms do not use hashtags at all, and their usefulness is limited. Third, using machine learning techniques such as topic modeling has serious limitations that should be considered (Chen et al. 2023). The classification relies on statistical dependencies and does not have to (often does not) correspond to researchers' theoretical assumptions, especially when complex phenomena are investigated (Grimmer & Stewart 2013). Supervised models (e.g., deep neural networks) may be more successful in detecting complex phenomena, but they are so-called black boxes (it is unknown why something is classified in a given way), and require time, effort as well as computational power to prepare training datasets (Di Franco & Santurro 2020). We suggest a method that is interpretable at its every step, requires less effort, time or computational power than supervised learning techniques, and quickly classifies large datasets.

The Keyword Algorithm

Our algorithm (Scheme 1) combines human involvement for advanced cognitive tasks, like classifying documents based on operational definitions, and automation for repetitive, computationally demanding tasks.

Scheme 1. The Keyword Algorithm

1. Specify a list of keywords K that are likely associated with the conspiracy theory topic.
2. Collect documents using K. As a result, a set of documents S is created.
3. Use word embeddings to find keywords similar to keywords $k \in K$ and likely associated with the conspiracy theory topic. Update the list K and repeat steps 1–3 until no new keywords appear.
4. Specify a list of keywords C that are likely associated with the way of thinking about the conspiracy theory topic.
5. Use word embeddings to find keywords similar to keywords $c \in C$ and likely associated with the conspiracy thinking. Update the list C and rerun steps 4–5 every time a new set of keywords are added to the existing list C.
6. Draw a random sample Q from the set S using keywords $c \in C$ as strata and a chosen equal stratum size (it should be large enough to make statistical inferences).
7. Manually classify documents in sample Q, that is decide according to definitions whether it contains conspiracy thinking or not.
8. Prepare a list of words (or ngrams of your choice) that occur in Q above a chosen frequency threshold (the threshold should be lower than the stratum size to include keywords $c \in C$ in computations). Use a manually annotated sample from dataset Q to calculate the probability of a document containing conspiracy thinking given that

it contains a keyword $c \in C$. The number of documents per word impacts uncertainty, therefore it is important to not only calculate probabilities but also to use a measure of uncertainty of your choice.

9. Based on the probability results, decide which words W_P indicate conspiracy thinking in the documents in Q and which indicate the opposite, i.e. W_N. Use decision thresholds of your choice. Verify the meaning of words to avoid false positive results (for instance, because of accidental co-occurrences).
10. Assign value +1 to words that indicate conspiracy thinking and −1 to words that indicate non-conspiracy thinking. Calculate the Conspiracy Thinking Index using the following formula: if (N of $W_P > 0$ and N of $W_N > 0$), then

$$CTI = -(N\ of\ W_N) \times \frac{N\ of\ W_p\ in\ a\ document}{\sqrt{N\ of\ words\ in\ a\ document}},$$

if (N of $W_P > 0$ and N of $W_N = 0$), then

$$CTI = \frac{N\ of\ W_p\ in\ a\ document}{\sqrt{N\ of\ words\ in\ a\ document}},$$

else $CTI = 0$.

11. Using manually coded dataset Q, verify which threshold of CTI gives the best evaluation metrics.
12. Using the CTI threshold from the previous step, classify all documents in S.
13. Repeat steps 1–13 every time new keyword(s) $k \in K$ are added to improve the list or new data are collected (take into account how quickly the studied concepts evolve).

The first three steps involve identifying a set of documents that may be relevant to the study topic. One common method is to use snowballing to find relevant keywords. We collected Twitter data using a list of words and hashtags related to COVID-19. To find new keywords (e.g. neologisms), we can create a table with the frequencies of words and look for other potentially relevant words. However, this approach can be challenging and error-prone due to the large number of out-of-context words that need to be manually verified. To overcome this issue, we propose using word embeddings on the collected documents.

Word embeddings are a technique that allows for the generation of vectors from the relationships between words. These vectors are thought to represent the meaning of words and make it possible to calculate the similarities between them. One of the significant advantages of word embeddings is that they capture complex relationships that may never appear together in the same document. For instance, the words ‘lawyer’ and ‘businessperson’ may never occur in the same sentence, but because they share a similar context (e.g., ‘suitcase,’ ‘money,’ ‘suit,’ ‘corporation’) they are considered as more similar to each other than, for instance, to ‘dancer’ (Kozlowski et al. 2019). We find this feature particularly useful in searching for conspiracy theories. Our procedure involves generating a network graph with the most similar words to our chosen keywords. By examining them in a given context, we can discover new conspiracy-related words. An updated list of keywords can then be used for the next iteration of data collection, generating new graphs and selecting other keywords. After several repetitions, a list of words likely associated with

the topic of conspiracy theories is produced. The significant advantage of this step is that the keywords include made-up words or ones whose original meaning has been changed by conspiracy theorists.

The first three steps result in a list of conspiracy-related words. The dataset is designed for recall to maximize relevant document inclusion. However, this may also lead to irrelevant documents. The next objective is to refine the dataset to include only documents that contain the target concept.

The fourth step involves creating a list of keywords that are likely associated with conspiracy thinking about the topic of conspiracy theories. These keywords will be used to detect conspiracy thinking within the specified general context, such as the pandemic. For example, the keyword 'gates' would return many unrelated results if used without context. However, when used in the dataset generated in the first step, it is likely to return results where Bill Gates is considered one of the designers of the COVID-19 pandemic. The process of selecting keywords for conspiracy thinking is the same as for collecting contextual data. The starting point is a list of keywords that indicate conspiracy thinking, such as words that mention main actors who conspire (e.g., Gates, Fauci, Luciferians, satanists), characteristic language that signals conspiracy (e.g., 'plandemic'), the aim of conspiracy (e.g., depopulation), or words characteristic of a specific conspiracy narrative (e.g., graphene oxide). Next, word embeddings can be used to iteratively collect a list of keywords that appear in the conspiracy context. This step can be partially automated, but it is necessary to manually select words based on human assessment to avoid false positive results and ensure high discrimination power.

The list created in the previous steps contains words that researchers believe may indicate conspiracy thinking. However, it is important to verify this through data analysis. Our algorithm recommends taking a random sample of data and manually coding them. Since the frequency of words is not equal in the dataset, simple random sampling can result in an imbalance of certain keywords. To address this, stratified sampling can be used with keywords as strata, resulting in a representation of documents for each keyword of interest. Additionally, determining the size of the strata allows us to control the minimum level of certainty. Due to the co-occurrence of some words, the sample size is often smaller than the number of words multiplied by the stratum size.

In the seventh step, human coding is required. It is suggested that the more complex the coding task is, the more important is to use at least more than one coder per document to achieve high-quality results (Matuszewski 2022; Grimmer & Stewart 2013).

Manually coding data enables researchers to pinpoint textual evidence of conspiracy thinking. The eighth step entails determining the probability of a document being classified as conspiracy thinking based on specific n-grams. While researchers can narrow down the list of keywords for which probabilities are calculated to those selected in the fifth step, it is recommended to analyze all n-grams in the dataset, including unigrams and bigrams, to uncover hidden patterns. Ultimately, the decision to consider an n-gram relevant lies with the researcher. In our study, we utilized Bayesian proportion tests to assess the probability of a document containing conspiracy theories if it includes a certain unigram or bigram. We used a liberal rule of choosing keywords with the 2.5% highest density interval above the probability of 0.55 as indicators of conspiracy thinking and keywords with the 97.5%

highest density interval below 0.1 as indicators of lack of conspiracy thinking. In other words, we identified keywords that suggest conspiracy theories and those that suggest the absence of conspiracy thinking. With these rules, we could identify keywords that likely indicate conspiracy theories as well as those that indicate the absence of conspiracy thinking, such as words that convey ridicule or mockery.

In Step 9, calculate the Conspiracy Thinking Index (CTI) for each document by assigning a value of 1 to conspiracy-related words and -1 to anti-conspiracy words based on the chosen thresholds. Begin by assigning these values to the selected keywords from the previous step.

In step 10, compute the Conspiracy Thinking Index . The rationale behind the formula is that specific phrases indicate conspiracy thinking, and the more of these phrases present in a document, the more likely it contains such thinking. However, there are no words that balance conspiracy thinking, making an arithmetic average informative, as in some basic sentiment analysis approaches. Conspiracy narrations can be used sarcastically, mocked, or ridiculed, and thus, there may be signals of it, such as certain phrases or emojis. However, these signals do not de-intensify conspiracy thinking, which would justify using the arithmetic average, but negate it. Thus, it is reflected in the formulas:

$$\text{if } N\ of\ W_N > 0, CTI = -(N\ of\ W_N) \times \frac{N\ of\ W_p\ in\ a\ document}{\sqrt{N\ of\ words\ in\ a\ document}} \text{ or}$$

$$\text{if } N\ of\ W_N = 0, CTI = \frac{N\ of\ W_p\ in\ a\ document}{\sqrt{N\ of\ words\ in\ a\ document}}.$$

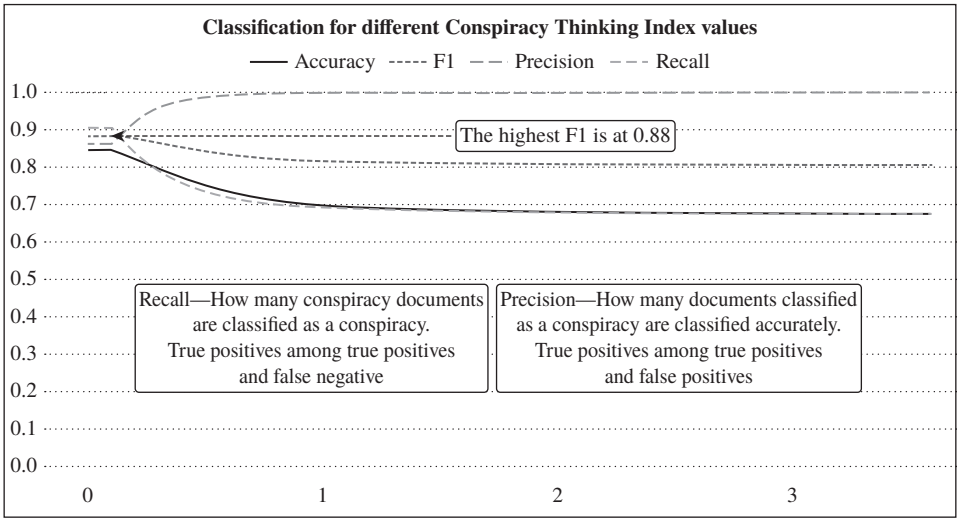
The fraction $\frac{N\ of\ W_p\ in\ a\ document}{\sqrt{N\ of\ words\ in\ a\ document}}$ divides the number of keywords signaling conspiracy thinking by the square root of a document's total number of words. Its purpose is to calculate the intensity of conspiracy thinking in a document. However, the more keywords signalizing negation, mockery, sarcasm etc. there are, the more likely it does not contain conspiracy thinking. Therefore, the index takes both negative and positive values, with $CTI < 0$ indicating that the respective document contains negative attitudes towards conspiracy theories or theorists, $CTI = 0$ indicating that the document does not contain conspiracy thinking, and $CTI > 0$ indicating that it does.

In the eleventh step, the manually coded dataset is utilized to calibrate the CTI based on a specific measure, such as recall, precision, or F1. The natural threshold of 0 can be used to classify all documents where CTI is greater than 0 as containing conspiracy thinking and those where CTI is less than or equal to 0 as not containing conspiracy thinking. However, the threshold can be adjusted to achieve specific outcomes. Increasing the threshold will improve precision but decrease recall, resulting in highly likely documents containing conspiracy thinking and excluding relevant documents (see Fig. 1).

After setting a threshold for CTI, the algorithm can classify all documents collected in steps 1–3. The process has an iterative nature, and it is recommended to repeat the entire process and update the results for new keywords based on the evolution of the studied narratives.

Figure 1

CTI threshold



Methods and Data

Survey study

According to the concept of the wisdom of crowds, collective decisions under certain conditions (diversity of opinions, independence, decentralization, synthesis) can be more accurate and reliable than decisions of individual experts (Surowiecki 2005). We used this idea to check if a group of people can generate a high-quality list of keywords detecting conspiracy theories related to the pandemic (RQ1). The study participants have different backgrounds (diversity condition), they express their opinion independently (independence condition), and based on their personal knowledge (decentralization condition). The answers were gathered using an online form (synthesis condition).

Our study included 163 participants, who were recruited from social science students at two Polish universities and members of a Facebook group dedicated to philosophy. The survey was conducted in April 2022, when the pandemic was still a significant topic of public discussion in Poland. We chose this timeframe to ensure that respondents could easily recognize and recall the concepts and related words we asked about. The public opinion seemed to be familiar with conspiracy theories regarding the coronavirus. According to a CBOS survey, in November 2020, 28% of adult Poles believed that the coronavirus pandemic was artificially triggered to reduce the human population living on Earth, and 45% that the pharmaceutical lobby, politicians and media around the world are deliberately exaggerating coronavirus risks (Cybulska & Pankowski 2020).

Eventually, we analyzed 154 responses (95% of total) and excluded cases where participants misunderstood the question or provided answers that were unworkable for our study (such as providing examples).

The respondents received a survey form to fill out with the following question:

> Imagine you have several million tweets from 2020–2022 containing words such as "coronavirus," "covid," "epidemic," "masks," and "vaccines." Please list words or phrases that come to your mind, the presence of which would indicate that the author of the tweet is a supporter of conspiracy theories related to the pandemic. Please separate the words/phrases with a comma or semicolon.

The respondents provided an average of 7.3 keywords, and in total, we gathered 690 unique keywords. Similarly to the results obtained by King et al. (2017), 82% of these words were mentioned by just one person, while only 11 words were recalled by more than ten people.

Social media dataset

We used data collected from Twitter using the academictwitteR package in R language (Barrie & Ho 2021). To ensure international comparability, we opted for a topic that is widely understood, namely conspiracy theories about the COVID-19 pandemic. The data was collected from November 1st, 2021, to January 31st, 2022, which happened to be a time of the highest COVID infections in Poland.

The dataset comprises 1,142,442 tweets, replies, and quotations in Polish. Ethical considerations necessitated the removal of all information besides text. Unique text strings were extracted and preprocessed by removing hyperlinks, Twitter mentions, punctuation, emojis, emoticons, and Polish stop words, and then converting the text to lowercase and lemmatizing it. Every tweet was considered a separate document.

Manual verification of CTI results

We used CTI to classify all documents in our dataset and drew a sample to manually verify if CTI correctly identified conspiracy theories. In order to have all the words selected by CTI in a sample for manual verification, we used stratified sampling. The aim of the sampling procedure was to collect at least 30 tweets for every word related (whether positively or negatively) to conspiracy. Tweets were drawn from the main dataset.

To account for potential cases missed by CTI, we intentionally add to the sample tweets that does not contain words selected by the index (20% of the sample size). Our sample size consists of 5710 documents, and four coders are responsible for classifying whether a tweet contains a pandemic-related conspiracy theory. Each tweet was coded twice, and the supervisor made the final decision when there was a discrepancy between the two codings.

Simulation of results and evaluation metrics

Comparisons between different keyword selection approaches may be affected by the random sampling and the division into training and testing datasets. Due to the randomness, the evaluation metrics are not constant, and it is crucial to ensure that the results are not outliers. To address this issue, the same sample size is used to calculate evaluation metrics of different methods, but this process is iterated 50 times. In other words, we simulate and compare results for 50 different samples.

Three evaluation metrics were used: precision, recall, and the F1 score. Precision informs about how many documents classified as containing conspiracy theories are in fact documents that contain conspiracy theories. Recall informs how many documents that contain conspiracy theory are actually classified as containing conspiracy theories. F1 score is a useful metric combining precision and recall as their harmonic mean. Therefore, we use the F1 score as our main dependent variable.

The two research questions require different setups.

To answer the first research question (RQ1: "Is the introduced method (CTI) a more accurate procedure of finding conspiracy theories than keyword search based on wisdom of crowds?") we compare classifications based on keywords selected by CTI with classifications based on keywords indicated by survey respondents. The number of keywords from the survey to be included can significantly impact the results. Rather than choosing a threshold arbitrarily, we calculated the results for words that were mentioned by at least 5, 10, 15, and 20 study participants.

The second research question (RQ2: "Is the introduced method (CTI) a more accurate procedure of finding conspiracy theories than supervised models?") compares CTI with four supervised learning models: convolutional neural network, XGBoost, support vector machines, and lasso classification. For every model, we used the *tidymodels* package in R.

It is widely accepted that larger training datasets lead to better model performance (Barberá et al. 2020). In our case, we believe this holds true as well. However, there is a drawback: larger datasets require more time, effort, and resources to manually code (Matuszewski 2022). Thus, we aim to find the most efficient method that delivers the best results while requiring the least resources. To do this, we refined our second research question and compared the performance of selected models and CTI for different sample sizes. We simulated 50 results each for samples with at least 10, 15, 20, 25, and 30 cases for every keyword with a CTI score above 0.

This study uses hierarchical linear regression models. The F1 scores are predicted by CTI and supervised models, but we specified sample size as a random effect group, because comparisons should be made between the methods that use the same number of datapoints. We used the Bayesian approach (rstanarm package).

Results

The keyword algorithm vs the wisdom of crowds

The first research question was about the difference between CTI and keyword search based on wisdom of crowds. For the comparison, we used the set of keywords selected by our algorithm and four sets of keywords indicated in the survey. Because no general rule exists regarding selecting a threshold of inclusion, we use keywords that appeared at least five, ten, fifteen, and twenty times and check how they affected the results.

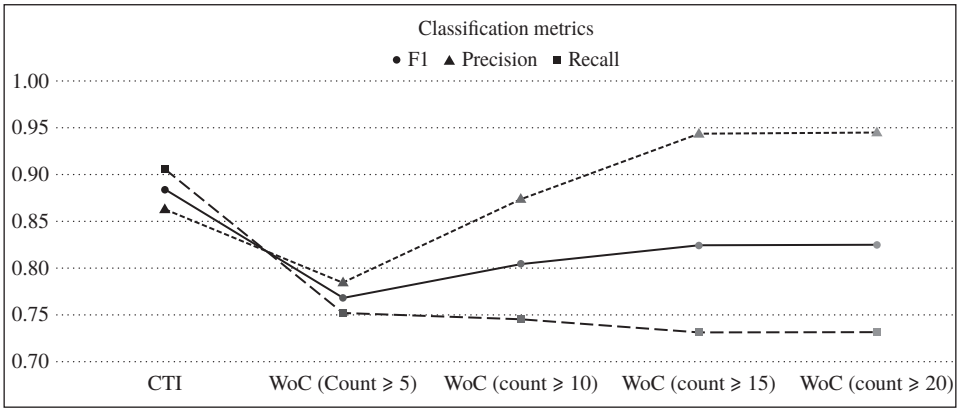
The F1 score using CTI is 0.88, with precision and recall close to this result at 0.86 and 0.91, respectively (see Fig. 2). For keywords that appear at least five times in the survey, the F1 score is 0.77, precision 0.79, and recall 0.75. When the threshold rises, the

keywords are more likely to identify conspiracy theories. We believe that it is congruent with the basic rules of the wisdom of crowds, i.e. that the more answers are given, the more precise the result is. However, this approach has a significant drawback. While it improves the identification of conspiracy theories, it also ignores relevant documents that should be classified as such. In our study, 27% of relevant documents were misclassified as not containing conspiracy theories when precision reached 0.95.

The results indicate that CTI generally surpasses WoC in terms of both precision and recall, with a higher F1 score of 0.88 for CTI compared to 0.82 for WoC. Although WoC may be more suitable for research focused solely on precision, it leads to biased datasets that fail to capture all aspects of the studied phenomenon. On the other hand, CTI consistently achieves higher recall, which is essential for detecting rapidly evolving phenomena.

Figure 2

Conspiracy Thinking Index vs Wisdom of Crowds



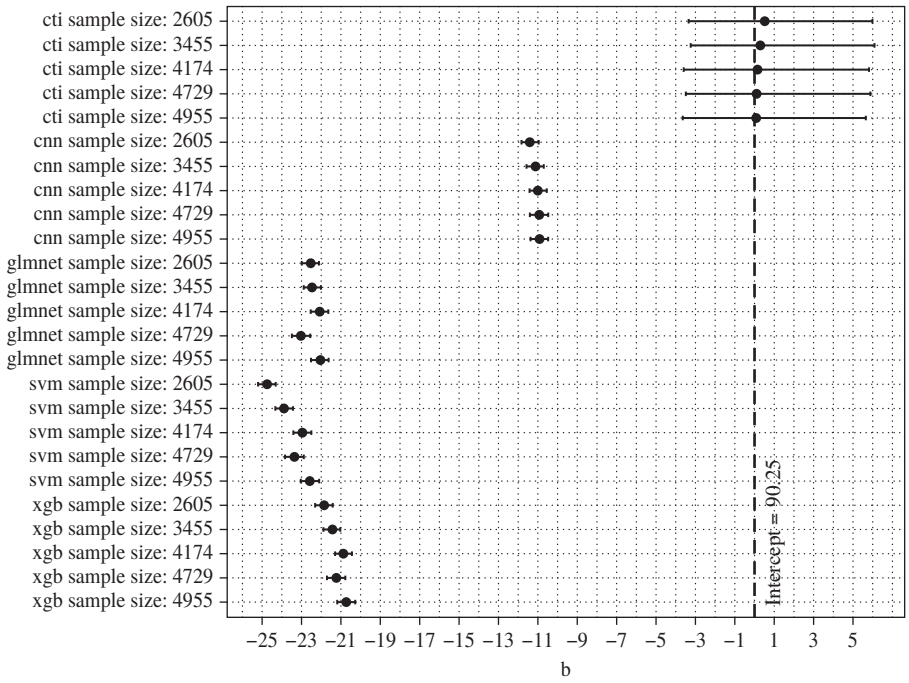
The Conspiracy Thinking Index vs Supervised Models

To answer the second research question, we compared the CTI's performance with popular supervised models. The mixed model's results show substantial and statistically significant differences between the CTI's and most supervised algorithms' evaluation statistics (see Fig. 3) The performance of CTI was not affected (95% HDI intervals did not overlap) by the number of manually coded documents (F1 score mean = 90.74 for sample size = 2,605; 90.43 for sample size = 3,455; 90.23 for sample size = 4,174; 90.17 for sample size = 4,729; 90.15 for sample size = 4,955). Such results were possible only for lasso classification when sample sizes exceeded 4,174 cases. The remaining classifiers performed significantly worse.

Overall, the effects for sample sizes appeared to be relatively low. It is noteworthy that the samples used for calculations are not random, but a product of the first phase of the algorithm. All of them consist of documents that are suspected to contain conspiracy, and at this level we compare whether the better approach would be to manually code the dataset and train supervised models or to omit human coding and use a much simpler method such

as CTI. The exact effects for the sample sizes depend on the model. As in the case of CTI, the differences are practically irrelevant. For convolutional neural networks, the increase of sample size is associated with a lower F1 score (mean = 82.4 for sample size = 2,605, and 80.9 for sample size = 4,955), for lasso classification it is associated with a higher F1 score (mean = 87.4 for sample size = 2,605, and 88.3 for sample size = 4,955), and for support vector machines and XGBoost the relationship is not statistically significant (when 95% HDI is used).

Figure 3



* Errorbars are 95% Highest Density Intervals

Discussion & Conclusions

Our proposed algorithm effectively combines human cognitive abilities and computer power to identify complex concepts within unstructured text data. We successfully tested its usefulness by applying it to the difficult case of conspiracy theories. Our algorithm generates a list of intuitive keywords, as well as those that are neologisms, have a different meaning from their official definition, or are unlikely to be considered by individuals not actively engaged in conspiracy discussions. Additionally, it can differentiate between conspiracy beliefs, discussions about them, and mockery. Furthermore, this method is capable of encompassing the diversity and changing nature of the searched phenomenon.

The algorithm identifies keywords and classifies the associated documents substantially better (in terms of F1 scores) than a group of 154 study participants. As in another study (King et al. 2017), people appeared to perform poorly at recalling many words, and to be heavily biased (the selection of words differed between people). Contrary to the results of King's et al., in our study they performed poorly at recalling a large number of words as well as finding words separating conspiracy thinking from other narrations. Even the most popular words were mentioned by 36%, meaning they were not considered obvious by the majority and could be omitted.

Our relatively simple formula for the Conspiracy Thinking Index also proved to be better (in terms of the F1 score), faster, and less computationally demanding than the selected supervised learning techniques.

Our research relates to the information retrieval literature regarding algorithms used to find keywords within search queries (Azcarraga et al. 2002; Carpineto & Romano 2012; Chen et al. 2009; Hristidis et al. 2008; Lin et al. 2017; Lu et al. 2014). However, it differs from it in several ways. First, we do not use synonyms or co-occurrences to expand the keyword list. We agree with Bai et al. (2005) that context-dependent query expansion would be a more useful approach. Advances in natural language processing enable new techniques, such as word embeddings (Matsui & Ferrara 2022), to find keywords meaningfully close to the ones used by conspiracy theorists. Such associations, visualized as a network, may be an approachable way for humans to verify which words should be checked as potentially good indicators of a specific notion. Second, our literature review on retrieving information and conspiracy theories on social media showed that some studies are focused on precision. For instance, they utterly rely on hashtags. Such an approach may significantly and artificially limit the studied phenomenon by omitting relevant data associated with unrecognized keywords. We address this issue by focusing both on precision and recall. Third, our study is not the first study trying to combine human and computer abilities to retrieve information. The most influential for our method and the reason to develop it was the study of King et al. (2017). Their idea was to use machine learning classifiers' mistakes to extract information that can be used in Boolean search strings. However, when we tried to implement this method, we found that it still required significant effort in creating a set of documents that would contain the searched notion. It should be large enough to make classifiers perform correctly, and the more complex a notion is, the more data they require. Most importantly, at least in the case of conspiracy theories, the classifiers did not provide useful information. Their mistakes were helpful in recognizing new keywords, but were still blind to the whole spectrum of other potentially relevant keywords.

Our aim was to produce an algorithm able to identify all relevant keywords that successfully distinguish between documents that contain and do not contain notions of interest. Our secondary aim was to make it as effortless and transparent as possible. While the first aim can be assessed objectively with the metrics we provide, the second one deserves a short discussion about limitations. Word embeddings require large amounts of textual data. The choice of new potential keywords is based on human intuition, meaning that expansion requires human work. However, the burden is the heaviest at the first iteration. In the following iterations, the new words may be highlighted on graphs, making them relatively easy to spot. The same applies to manual coding of documents. The number

of documents to code depends on the number of new keywords, therefore the workload is the highest at the beginning, where all keywords need to be checked. The results of our experiments show that due to co-occurrences of words, even ten documents may be enough to make confident inferences.

Our proposed algorithm performed exceptionally well in classifying complex notions in documents. It is fast and requires minimal human coding, making it a more efficient alternative to supervised techniques. Additionally, it is transparent, with the keyword list relying heavily on human intuition guided by statistical results. The Conspiracy Thinking Index formula, which forms the basis of the classification, is straightforward and easy to understand. The algorithm is iterative, meaning it updates the list of keywords and only requires a small workload after the first iteration. It is also adjustable, making it suitable for evaluating other complex notions, even those that are theoretically defined, evolving, or difficult to detect due to censorship or platform algorithms.

Funding

This research was funded by National Science Centre, Poland (grant no 2020/39/I/HS5/00176).

References

Ahmed, W., Vidal-Alaball, J., Downing, J., & Seguí, F. L. 2020. COVID-19 and the 5G conspiracy theory: Social network analysis of Twitter data, *Journal of Medical Internet Research 22*(5): e19458.

Azcarraga, A. P., Yap, T., & Chua, T. S. 2002. Comparing keyword extraction techniques for websom text archives, *International Journal on Artificial Intelligence Tools 11*(02): 219–232. https://doi.org/10.1142/S0218213002000861

Bai, J., Song, D., Bruza, P., Nie, J.-Y., & Cao, G. 2005. Query expansion using term relationships in language models for information retrieval. *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, pp. 688–695. https://doi.org/10.1145/1099554.1099725

Barberá, P., Boydstun, A. E., Linn, S., McMahon, R., & Nagler, J. 2020. Automated Text Classification of News Articles: A Practical Guide, *Political Analysis 29*(1): 1–24. https://doi.org/10.1017/pan.2020.8

Barrie, C., & Ho, J. C. 2021. academictwitteR: An R package to access the Twitter Academic Research Product Track v2 API endpoint, *Journal of Open Source Software 6*(62): 3272. https://doi.org/10.21105/joss.03272

Bessi, A., Coletto, M., Davidescu, G. A., Scala, A., Caldarelli, G., & Quattrociocchi, W. 2015. Science vs conspiracy: Collective narratives in the age of misinformation, *PloS One 10*(2): e0118093.

Carpineto, C., & Romano, G. 2012. A Survey of Automatic Query Expansion in Information Retrieval, *ACM Computing Surveys 44*(1): 1:50. https://doi.org/10.1145/2071389.2071390

Chen, Y., Peng, Z., Kim, S.-H., & Choi, C. W. 2023. What We Can Do and Cannot Do with Topic Modeling: A Systematic Review, *Communication Methods and Measures 17*(2): 1–20. https://doi.org/10.1080/19312458.2023.2167965

Chen, Y., Wang, W., Liu, Z., & Lin, X. 2009. Keyword search on structured and semi-structured data, *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data*, pp. 1005–1010. https://doi.org/10.1145/1559845.1559966

Cybulska, A., & Pankowski, K. 2020. *Koronasceptyzm, czyli kto nie wierzy w zagrożenie epidemią* (158/2020). CBOS. https://www.cbos.pl/SPISKOM.POL/2020/K_158_20.PDF

Czech, F. 2019. Saturation of the media with conspiracy narratives: Content analysis of selected Polish news magazines, *Środkowoeuropejskie Studia Polityczne 2*: 151–171.

Di Franco, G., & Santurro, M. 2020. Machine learning, artificial neural networks and social research, *Quality & Quantity 55*(3): 1007–1025. https://doi.org/10.1007/s11135-020-01037-y

DiFonzo, N. 2019. Conspiracy Rumor Psychology, in: J. E. Uscinski (ed.), *Conspiracy Theories and the People Who Believe Them* (pp. 257–268). Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780190844073.003.0017

Furnas, G. W., Landauer, T. K., Gomez, L. M., & Dumais, S. T. 1987. The vocabulary problem in human-system communication, *Communications of the ACM 30*(11): 964–971. https://doi.org/10.1145/32206.32212

Giachanou, A., Ghanem, B., & Rosso, P. 2021. Detection of conspiracy propagators using psycho-linguistic characteristics, *Journal of Information Science*, 0165551520985486.

Glowacki, E. M., & Taylor, M. A. 2020. Health hyperbolism: A study in health crisis rhetoric, *Qualitative Health Research 30*(12): 1953–1964.

Goreis, A., & Voracek, M. 2019. A Systematic Review and Meta-Analysis of Psychological Research on Conspiracy Beliefs: Field Characteristics, Measurement Instruments, and Associations With Personality Traits, *Frontiers in Psychology 10*. https://www.frontiersin.org/article/10.3389/fpsyg.2019.00205

Grimmer, J., & Stewart, B. M. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts, *Political Analysis 21*(3): 267–297.

Guzek, D. 2021. When partisan groups get access to the digital society: Re-voicing religion in Poland, *Information, Communication & Society 26*(6): 1–16. https://doi.org/10.1080/1369118X.2021.1994628

Havey, N. F. 2020. Partisan public health: How does political ideology influence support for COVID-19 related misinformation?, *Journal of Computational Social Science 3*(2): 319–342.

Hayes, P., & Weinstein, S. P. 1990. CONSTRUE/TIS: A System for Content-Based Indexing of a Database of News Stories. *IAAI*.

Hegel, G. W. F. 1977. *Phenomenology of Spirit* (A. V. Miller, Trans.). Oxford: Oxford University Press.

Hofstadter, R. 1996. *The Paranoid Style in American Politics: And Other Essays* (First Edition Used). Cambridge, MA.: Harvard University Press.

Houli, D., Radford, M. L., & Singh, V. K. 2021. "COVID19 is_": The Perpetuation of Coronavirus Conspiracy Theories via Google Autocomplete, *Proceedings of the Association for Information Science and Technology 5*(1): 218–229.

Hristidis, V., Hwang, H., & Papakonstantinou, Y. 2008. Authority-based keyword search in databases, *ACM Transactions on Database Systems 33*(1): 1–40. https://doi.org/10.1145/1331904.1331905

Kant, G., Wiebelt, L., Weisser, C., Kis-Katos, K., Luber, M., & Säfken, B. 2022. An iterative topic model filtering framework for short and noisy user-generated data: Analyzing conspiracy theories on twitter, *International Journal of Data Science and Analytics*: 1–21.

King, G., Lam, P., & Roberts, M. E. 2017. Computer-Assisted Keyword and Document Set Discovery from Unstructured Text, *American Journal of Political Science 61*(4): 971–988.

Kou, Y., Gui, X., Chen, Y., & Pine, K. 2017. Conspiracy talk on social media: Collective sensemaking during a public health crisis, *Proceedings of the ACM on Human-Computer Interaction* 1: 1–21.

Kozlowski, A. C., Taddy, M., & Evans, J. A. 2019. The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings, *American Sociological Review 84*(5): 905–949. https://doi.org/10.1177/0003122419877135

Lewandowsky, S., & Cook, J. 2020. *The Conspiracy Theory Handbook*. https://www.climatechangecommunication.org/all/handbook/the-conspiracy-theory-handbook/

Lin, C., Wang, J., & Rong, C. 2017. Towards heterogeneous keyword search, *Proceedings of the ACM Turing 50th Celebration Conference—China*: 1–6. https://doi.org/10.1145/3063955.3064802

Lu, Y., Lu, J., Cong, G., Wu, W., & Shahabi, C. 2014. Efficient Algorithms and Cost Models for Reverse Spatial-Keyword k-Nearest Neighbor Search, *ACM Transactions on Database Systems 39*(2): 13:1–13:46. https://doi.org/10.1145/2576232

Łukowski, P. 2016. Sens wyrażenia „teoria spiskowa," czyli jak odróżnić teorię spiskową od niespiskowej, *Internetowy Magazyn Filozoficzny Hybris* 33: 101–124.

Mahl, D., Schäfer, M. S., & Zeng, J. 2023. Conspiracy theories in online environments: An interdisciplinary literature review and agenda for future research, *New Media & Society 27*(7). https://doi.org/10.1177/14614448221075759

Marcellino, W., Helmus, T. C., Kerrigan, J., Reininger, H., Karimov, R. I., & Lawrence, R. A. 2021. *Detecting Conspiracy Theories on Social Media: Improving Machine Learning to Detect and Understand Online Conspiracy Theories*. RAND Corporation. https://doi.org/10.7249/RR-A676-1

Matsui, A., & Ferrara, E. 2022. *Word Embedding for Social Sciences: An Interdisciplinary Survey* (arXiv:2207.03086). arXiv. http://arxiv.org/abs/2207.03086

Matuszewski, P. 2022. How to prepare data for the automatic classification of politically related beliefs expressed on Twitter? The consequences of researchers' decisions on the number of coders, the algorithm learning procedure, and the pre-processing steps on the performance of supervised models, *Quality & Quantity 57*(1): 301–321. https://doi.org/10.1007/s11135-022-01372-2

Napolitano, M. G., & Reuter, K. 2021. What is a conspiracy theory?, *Erkenntnis 88*: 1–28.

Nugier, A., Limousi, F., & Lydié, N. 2018. Vaccine criticism: Presence and arguments on French-speaking websites, *Medecine et Maladies Infectieuses 48*(1): 37–43.

Pilch, I., Turska-Kawa, A., Wardawy, P., Olszanecka-Marmola, A., & Smołkowska-Jędo, W. 2023. Contemporary trends in psychological research on conspiracy beliefs. A systematic review, *Frontiers in Psychology 14*. https://doi.org/10.3389/fpsyg.2023.1075779

Poberezhskaya, M. 2018. Blogging about climate change in Russia: Activism, scepticism and conspiracies, *Environmental Communication 12*(7): 942–955.

Popper, K. R. 2013. *The Open Society and its Enemies*. Princeton: Princeton University Press.

Samory, M., & Mitra, T. 2018. 'The Government Spies Using Our Webcams' The Language of Conspiracy Theories in Online Discussions, *Proceedings of the ACM on Human-Computer Interaction 2*: 1–24.

Seawright, J., & Gerring, J. 2008. Case Selection Techniques in Case Study Research: A Menu of Qualitative and Quantitative Options, *Political Research Quarterly 61*(2): 294–308. https://doi.org/10.1177/1065912907313077

Shahsavari, S., Holur, P., Wang, T., Tangherlini, T. R., & Roychowdhury, V. 2020. Conspiracy in the time of corona: Automatic detection of emerging COVID-19 conspiracy theories in social media and the news, *Journal of Computational Social Science 3*(2): 279–317.

Stasielowicz, L. 2022. Who believes in conspiracy theories? A *meta*-analysis on personality correlates, *Journal of Research in Personality 98*. https://doi.org/10.1016/j.jrp.2022.104229

Surowiecki, J. 2005. *The Wisdom of Crowds* (Reprint edition). Anchor.

Tangherlini, T. R., Shahsavari, S., Shahbazi, B., Ebrahimzadeh, E., & Roychowdhury, V. 2020. An automated pipeline for the discovery of conspiracy and conspiracy theory narrative frameworks: Bridgegate, Pizzagate and storytelling on the web, *PloS One 15*(6): e0233879.

Tyagi, A., & Carley, K. M. 2021. Climate Change Conspiracy Theories on Social Media, *arXiv Preprint arXiv:2107.03318*.

Uscinski, J. E. 2020. *Conspiracy Theories: A Primer*. Rowman & Littlefield Publishers.

Vermeule, C. A., & Sunstein, C. R. 2009. Conspiracy Theories: Causes and Cures, *Journal of Political Philosophy 17*(2): 202–227.

Walker, J. 2018. What We Mean When We Say "Conspiracy Theory," in: J. E. Uscinski (ed.), *Conspiracy Theories and The People Who Believe Them* (pp. 53–61). Oxford University Press.

Walker, J. 2019. What We Mean When We Say "Conspiracy Theory," in: J. E. Uscinski (ed.), *Conspiracy Theories and the People Who Believe Them* (pp. 53–61). Oxford University Press. https://doi.org/10.1093/oso/9780190844073.003.0003

Zonis, M., & Joseph, C. M. 1994. Conspiracy thinking in the Middle East, *Political Psychology 15*(3): 443–459.

Biographical Notes:

Paweł Matuszewski (Ph.D. hab.) is an associate professor in sociology at Collegium Civitas (Warsaw, Poland). He has expertise in public opinion, political communication, political behavior, social media, digital behavior and opinion mining. In his work, he focuses on the micro foundations of macro-level phenomena.

ORCID iD: 0000-0003-0069-157X

E-mail: pawel.m.matuszewski@gmail.com

Michał Rams-Ługowski is a PhD student at the Doctoral School of the University of Silesia in Katowice (Poland). He works on the application of philosophy and political economy in Communication & Media Studies. His main interest lies in researching media as means of managing the commons and the production of subjectivity.

ORCID iD: 0000-0002-5815-0875

E-mail: michal.rams-lugowski@us.edu.pl

Appendix
Top 50 most frequent words