Caught in a networked collusion? Homogeneity in conspiracy-related discussion networks on YouTube

1. Introduction

The potential threat to democratic societies of widespread misinformation in the form of conspiracy beliefs has previously been the subject of public discussions [1], [2]. Research has referred to conspiracy beliefs as narratives about secret and powerful forces following plots that harm certain groups of society and benefit those forces [3]. Examples are the beliefs that the moon landing was faked by NASA, that the CIA is responsible for the John F. Kennedy assassination, or that vapor trails from airplanes (so-called Chemtrails) are sprayed by governments to manipulate the population's health [4], [5]. The dissemination of conspiracy beliefs poses a hazard to individuals and societies, since exposure to material promoting conspiracy ideation decreases recipients' intentions to engage in politics [6] and pro-social activities [7]. It can also have an impact on political decisions (e.g., voting [6]) and health decisions (e.g., intention to vaccinate [8]).

Social media services such as YouTube appear to be ideal venues to circulate conspiracy beliefs among the population and insinuate a public sentiment that resonates with those conspiratorial beliefs. While there is initial evidence about the circulation of conspiracy theories in social media networks, less is known about the particular sub-networks in which conspiracist content is discussed and how these online networks are structured.

Given that conspiracy theorists (people who believe or formulate conspiracy theories) often represent minorities that face the risk of being segregated from society [9], [10], it appears crucial to map the composition and interconnections of the online networks that promote such theories. Evidence about the network structure in those topical contexts could help to address the question of to what extent social media communication enables users to become caught in like-minded, that is, homogeneous clusters without exposure to cross-cutting views [11], [12].

Corresponding Author: Daniel Röchert, University of Duisburg-Essen, Department of Information Science and Applied Cognitive Science, Group Digital Citizenship in Network Technologies; Forsthausweg 2, 47057 Duisburg. Email: daniel.roechert@uni-due.de

Addressing the prevalence of homogeneity in online networks, previous research indicated that in the case of networks discussing three politically relevant topics on YouTube, follow-up comments were more likely to express opposing views than similar opinions [13]. Discussions on conspiracy theories, however, might be structured differently: The spiral of silence theory [14] predicts that people are more inclined to express themselves in situations in which they feel part of the majority. Drawing on this, one could assume that supporters of conspiracy theories—as social minorities—might only voice their viewpoints in contexts in which they encounter agreement. At the same time, believing in a conspiracy theory often goes hand in hand with a need for uniqueness, that is, the wish to stand out from the mass [15], which could also lead supporters to interact with opponents of conspiracy theories. These two lines of reasoning do not really allow a prediction about the level of homogeneity within networks in which conspiracy theories are discussed: Do people only interact with each other when they agree that the conspiracy theory is valid (or true)? Addressing this question will contribute to identifying whether there are specific groups in societies that are more susceptible to becoming caught in homogeneous communication clusters filled with like-minded views.

The issue of conspiracy theories on social media has received a lot of critical attention, especially in times of the current COVID-19 pandemic, where false news is spreading on different online social media [16]. One of the greatest challenges here is correctly identifying content that contains and supports misinformation and spreads it in the form of videos and comments on social media. Using big data and machine learning methods, models can be trained to predict this content with supporting conspiracy theory content. Currently, there is no data available that links conspiracy-theory videos with conspiracy-theory comments and additionally computes their opinion-based homogeneity to be able to express conclusions about their relationships and communication pattern.

By employing natural language processing (NLP) and social network analysis, this research is intended to examine the presence of conspiracy theories and associated discussion networks on the video sharing platform YouTube. More specifically, this study analyzes: (a) The prevalence of videos that promote or debunk conspiracy theories on YouTube; (b) the social context, that is, user-generated comments, likes, and dislikes, which accompanies videos on conspiracy theories; and (c) the interconnection of discussion networks associated with those videos, that is, the opinion-based homogeneity among user-generated comments.

The paper is organized as follows: In Section 2, we explain the theoretical background of conspiracy theories in social media and their relation to the spiral of silence of homogeneous/heterogeneous groups. We present our research method consisting of the description of the dataset, the annotation of the data, the machine learning model BERT, and the network analysis in Section 3. Section 4 summarizes the results of our study and discusses them in Section 5. Finally, in Section 6, we conclude with a summary of findings and future research potential.

2. Theoretical background

2.1. Conspiracy theories in social media

The spread of misinformation in the digital age is a pressing problem that has been researched on different social media platforms such as Twitter [17], [18], Facebook [2], [19], and YouTube [20], [21], focusing on different topics (e.g., vaccinations, rumors, and conspiracy theories). The rapid increase in users on online social networks such as YouTube or Facebook and their published content also poses risks for other users, for instance, in the form of false information, cyber bullying, and pornographic material [22]. A report by the Reuters Institute in 2019 showed that the presence and spread of misinformation are perceived—globally—as an urgent problem, especially when it comes to trusting platforms that post public content. In a survey covering 38 countries, 55% participants of all countries

are concerned about not being able to distinguish between what is real and what is fake on the Internet. More specifically, 85% of Brazilians, 70% of Britons, and 67% of Americans worry about what is real and what is fake on the Internet. In Germany (38%) and the Netherlands (31%), the prevalence of concerns is lower [23]. Despite this mistrust in online platforms, the number of users following local news on the Internet is growing. According to a recent study, 89% of news is retrieved digitally, which includes news websites, apps, and social media [24].

Misinformation spread through different online channels can manifest itself in the form of conspiracy theories. While a variety of definitions of the term 'conspiracy theory' have been suggested, this work relies on the definition suggested by Keeley [68, p. 116], who referred to a conspiracy theory as a “proposed explanation of some historical event (or events) in terms of the significant causal agency of a relatively small group of persons—the conspirators—acting in secret.”

With the emergence of social media channels allowing the broadcast of user-generated content and citizens’ responses associated with that content, research has offered initial evidence on the role conspiracy beliefs are playing in those communication environments [26]. To analyze this content, software tools are needed; tools like HarVis provide a way to get more information on specific topics and also perform different analytics to capture, process, and visualize social media content on YouTube [27]. Focusing on the prevalence of conspiracy beliefs, Bessi and colleagues found that dealing with conspiracy pages on YouTube and Facebook goes hand in hand with users’ polarization and their presence in homogeneous communication networks [1], [2], [28]. However, exposure to conspiracy beliefs can occur even without being polarized or deliberately seeking information on conspiracies: Allgaier [29] revealed that small variations of search terms on YouTube can lead users who are interested in science either to scientific videos or to material that promotes conspiracy beliefs framed as serious scientific evidence, making it more difficult for users to differentiate between truths and falsehoods. Another study in the context of the Zika virus

outbreak in 2016 demonstrated that 12 out of 35 videos related to that topic contained conspiracy beliefs [30]. The numbers of user responses (i.e., comments, replies, likes), however, did not differ between conspiracy and non-conspiracy videos. The researchers Wood & Douglas [26] analyzed comments from several news articles on the conspiracy theory of 9/11 and found that authors of conspiratorial comments are more inclined to believe other unrelated conspiracy theories, which is in line with past research [31], [32]. Furthermore, the analysis of 1,459 conspiracy-supportive comments indicated a higher level of mistrust expressed than in anti-conspiracy-theory comments. Responding to the presence of conspiracy theories expressed in articles, videos, and user-generated comments in social media, there are attempts to counterargue or even explicitly debunk this kind of misinformation by creating and spreading so-called counter-messages [33]–[35]. An experiment showed that the rectification of misinformation in social media by presenting counter-arguments can be successful and reduce the amount of misperceptions [36]. Another study yielded mixed results, revealing that messages rejecting a conspiracy theory (e.g., on vaccination) and responding with counter-arguments can succeed, but only if these are present prior to the arguments of the conspiracy theories [37]. While counter-messages and corrective information might help to combat conspiracy theories online, Weeks & Gil de Zúñiga [35] argue that users do not commonly encounter these counter-messages "in the wild." Research on counter-messages as responses to extremist videos on YouTube, however, showed that due to the YouTube recommendation algorithm, counter-messages are directly associated with extremist videos. Therefore, those users who view counter-videos are likely to receive a recommendation for extremist videos (which the counter-video intended to debunk in the first place) [34].

2.2. User reactions as an influential social context of conspiracy theories in social media

While the presence and spread of conspiracy theories in social media represent a societal problem per se, a comprehensive analysis needs to examine the social context in

which conspiracist content is embedded [35]. As pointed out by previous research, the characteristic nature of social media is that news articles, status updates, tweets, and videos are integrated in a social context that can be made up of different types of user reactions such as likes, dislikes, or user-generated comments [38]. This line of research also argued that this social context in particular has the potential to either undo, that is, weaken, or reinforce the effects of the original content (e.g., status update or video). For instance, a YouTube video promoting the "Pizzagate" conspiracy (e.g., the debunked theory that high-ranking U.S. political officials led a global child-trafficking ring using a Pizza restaurant in Washington D.C. as their headquarters) may have a greater chance of persuading viewers when it is accompanied by a high number of likes and user-generated comments claiming that they knew that Hillary Clinton's campaign was corrupt and involved in dubious businesses.

According to the bandwagon heuristic, human beings rely on information in their environment that could reflect that a majority of, or at least many other, people agree with a certain claim and, therefore, this claim must be right (following the rule "what is popular must be good") [39], [40]. Numeric information in terms of a high number of likes, thus, could serve as an important indicator for individuals, reflecting that many others approved this message, which, consequently, must be valid.

Likewise, exemplification theory [41] suggests that vivid examples of a complex issue are easier to process psychologically and, therefore, have a greater chance of affecting individuals' judgments. User-generated comments are intended to serve as these kinds of examples, concretely representing a certain stance (e.g., a pro-conspiracy theory viewpoint: "I truly believe that this child sex ring exists") or personal experience: ("I've read the Clinton emails and they clearly show this ring exists") and could be used by readers or viewers as a basis for estimates about how society might also think about this issue [42]. Following these theoretical considerations, a higher number of likes or user-generated comments associated with videos on conspiracy theories could either fortify or attenuate the persuasive effects of

the original video. Given these potential effects, we are interested in examining the extent to which user reactions (number of likes, dislikes, and comments) vary between YouTube videos that promote versus challenge a conspiracy theory (Research Question 1).

However, not only the amount of user reactions might be indicative of how the social context qualifies effects of the video, but also the actual content of those reactions. While numeric information in terms of views, likes, and dislikes are perceived by users as ambiguous cues [43], a series of studies has shown that the valence of user-generated comments (e.g., supporting versus opposing a theory) as vivid exemplifications of experiences or opinions can either shape the evaluation of the original message (e.g., the YouTube video) [44], influence readers' or viewers' personal attitudes [45], or affect the opinion climate that recipients project onto the general population [42]. Against this background, we ask which opinion climate is reflected in user-generated comments (i.e., the distribution of supporting comments versus opposing comments) associated with YouTube videos on conspiracy theories (Research Question 2).

2.3. Believers in conspiracy theories: A minority in society

From a societal point of view, people who believe or express support for conspiracy theories can commonly be seen as minorities and, in many cases, as marginalized groups [9], [10]. This marginalized position in society may have implications for the communication behavior of people with conspiracy beliefs. The spiral of silence theory [14] suggests that human beings are driven by the wish to be accepted by their social environment. Pursuing this goal of social approval, they feel comfortable with expressing viewpoints when they are in line with the prevailing opinion climate around them. At the same time, they withhold their personal stance when they realize that this viewpoint deviates from the mainstream, or at least from the opinion trend around them. Consequently, one could assume that individuals with conspiracy beliefs feel comfortable when discussing their views with others who also believe

in the same theory and, at the same time, they avoid interactions with those who offer challenging views and could reject them for thinking differently. In the long run, this communication pattern could lead marginalized minorities to, themselves, be caught in comfortable, like-minded cocoons that solely confirm their worldview and represent a segregated cluster of homogeneous information and discussion.

Indeed, this formation of homogeneous subgroups in online communication has been a longstanding concern since the emergence of the Internet and even more salient since the rise of social media platforms [11], [12]. Empirical evidence, however, has repeatedly shown that while social media users are more likely to be connected with like-minded others, they—often incidentally—interact and become exposed to content or opinionated messages that challenge their personal viewpoint or ideology [13], [46]–[50].

The conclusion of this line of research is that especially those individuals who are politically extreme or at the margins of society are more likely to interact in homogeneous communication spaces. People that believe in conspiracy theories, however, may not always be politically extreme or members of marginalized groups [51]; therefore, it is unclear to what extent networks in which conspiracy theories are discussed are homogeneous in terms of the opinions expressed therein. Theoretically, different communication structures among conspiracy theory believers and non-believers are conceivable. Based on the spiral of silence theory [14], one could argue that those who support the validity of conspiracy theories—as a minority—would only interact with those who think alike. Research on minorities and their potential influence could challenge this view: A recent study revealed that a conspiracy mentality is often driven by the wish to stand out from the crowd and feel unique in contrast to the majority [15]. Thus, the status of being in a unique minority could be a driver to seek encounters with majority members who challenge one's conspiracy views. This, in fact, could even prove to be effective, given that minority influence research has suggested that minorities are able to modify mainstream ideas or opinions and persuade majority members

by expressing their viewpoint consistently across time and situations [52], [53]. If social media users with conspiracy beliefs are driven by the motive to change the opinion landscape and inject their beliefs or theories into the mainstream, it seems likely that they are going to interact with users holding and expressing diverging views in order to persuade them.

Opinion-based homogeneity is a concept that can assess the structure of discussion networks and refers to the extent to which opponents, such as believers and non-believers of conspiracy theories, might be interconnected. Specifically, this concept refers to the degree to which messages (e.g., user-generated comments) that are semantically similar are connected in the network. Opinion-based homogeneity [13] can be measured by the global E-I index [54], which is defined as follows:

$$EI\ Index = \frac{\mathrm{E} - \mathrm{I}}{\mathrm{E} + \mathrm{I}}$$

where E is the number of external ties (ties between users and videos with different stances) and I is the number of internal ties (ties between users and videos with the same stance). The interpretation of the resulting index ranges from –1.0 to +1.0. In a completely heterogeneous network, a value of +1.0 indicates that there are no links between nodes of the same group, while in a homogeneous network, a value of –1.0 indicates that all links between nodes are connected to their specific group. A value of 0 indicates that the ties occur equally often. Additionally, to obtain a more precise picture of the homogeneity in the network, the calculation can also be performed individually for each class to compare the respective opinions in the network.

Applying the concept of opinion-based homogeneity and its operationalization to discussions of conspiracy theories on YouTube, we ask to what extent users with a core opinion for or against a conspiracy engage with the different types of videos (Research Question 3).

3. Method

In the following, we first describe the dataset that forms the basis of the analysis. We then outline two different methods for annotations: One to assess the valence of the videos and the other to identify the valence of user-generated comments. The state-of-the-art NLP deep learning model BERT (Bidirectional Encoder Representations from Transformers) was then trained with the annotated data to classify the remaining comments. Furthermore, we compared the performance of BERT with other machine learning baseline models (i.e., Logistic Regression (LR) and Support Vector Machine (SVM)), where BERT showed the best results. Using network analysis, the annotated videos and the predicted comments were linked together to build a network structure representing the discussion landscape of published videos. By measuring opinion-based homogeneity, it is possible to understand the overall role of homogeneous versus heterogeneous communication ties in the network between individual users and videos.

3.1. Dataset

To examine the presence and discussion networks of conspiracy theories on YouTube, we focused on three classic conspiracy theories that had already been investigated in previous research [55]–[57]. On December 22, 2018, we performed an automatic data collection on YouTube using the terms “Chemtrails Conspiracy,” “Hollow Earth Conspiracy,” and “New World Order Conspiracy” and filtered the query for English language content, sorted by the number of views. To assure that data collection explicitly deals with conspiracy theories, we decided to always include the word “conspiracy” in the YouTube search. Without this addition, content that was not related to the actual conspiracy theory would incorrectly be displayed. This is the case, for instance, with the search term “New World Order,” where the most frequent hits are music videos by the band “New Order.” We decided to focus on the videos that were viewed the most and, therefore, deliberately stopped data collection at about 100 records per topic. Furthermore, we considered the number of views, likes, or comments

as an indicator of the extent to which the conspiracy theory had encountered attention on YouTube. The search term for the conspiracy theory of Hollow Earth, however, did not yield more than 89 hits. Table 1 provides an overview of the crawled videos with their corresponding search terms and the data provided by this crawling.

Table 1. Overview of YouTube data. *Note: Videos in the category "neither" are not included in the table and some videos were deleted from YouTube; therefore, the distribution of conspiracy videos is unequal*.

Conspiracy theory	Videos	Views	Likes	Dislikes	Comments	Channels
Hollow Earth	59	8,630,996	65,686	10,521	24,146	51
Chemtrails	61	14,877,499	321,098	24,868	122,074	57
New World Order	56	16,504,447	168,084	13,836	40,717	49

Further information on the core concepts of investigated conspiracy theories can be found in Online Appendix A.

3.2. Annotation

In this study, the annotation contains two essential components: First, the coding of the video material and its content in order to obtain information on the stance of the video, and second, the annotation of the comments and replies on these selected videos. In the following two sections, we outline the procedure in greater detail.

3.2.1. Manual video labeling

Given that it is not always possible to infer whether a video advocates in favor or against the validity of a conspiracy theory based on metadata (e.g., the title of the video), all videos were examined by three annotators according to the following classes: supporting theory, debunking theory, or neither. For the classification of the videos, a total of 75 videos per conspiracy theory were labeled (some videos have been removed from the analysis since YouTube deleted them). For the classification of the videos, the title and description of the video were considered. The minimum video length in the dataset was 37 seconds, while the maximum duration of a video was up to 2:29 hours.

To measure the reliability of our labels, we created a smaller, likewise randomized, dataset of 40 of the 75 videos per conspiracy theory, which was then labeled by a fourth annotator. The overall percentage agreement for the Hollow Earth dataset was 50%, while the New World Order dataset reached 72.5%. The Chemtrail dataset had the highest value with an agreement of 80%. Since the reliability of the Hollow Earth dataset was comparatively low, two annotators and the first author of the paper independently went through all videos which yielded disagreement. Meanwhile, notes were documented for each video, on the basis of which the decision was justified. The preferred classes were then expressed one after the other, and, in the event of disagreement, the first author's decision was added to obtain a majority decision. Through the second round of evaluation, we were able to assign a distinct class and improve the quality of the annotation. We did not use this procedure for the other two conspiracy theories, since the intercoder agreement was satisfactory. Accordingly, we kept the classes of the first annotation for the other two datasets. The annotation of the videos is important due to the fact that it can be linked to the opinion-based homogeneity of the comments and replies to identify homogeneous spaces in the network.

3.2.2. Manual comments labeling

For the annotation of the YouTube comments, we chose the crowd-sourcing platform Amazon Mechanical Turk to annotate 8,000 randomly selected comments and replies per conspiracy theory. Comments were categorized into one of three classes (pro-theory, contra-theory, other). Online Appendix B contains the complete coding scheme of the comments and a detailed description of the classes.

For the annotation of the data for each conspiracy theory, all annotators received rules and examples for coding the comments (see Online Appendix B). To increase the quality of the collected data, we set the Human Intelligence Task (HIT) Approval Rate (%) for all requesters' HITs greater than 95 and the number of HITs approved greater than 5,000. For

each comment we paid $0.01 to the annotators. To take the reliability of comment annotation into account, each comment was annotated by three annotators.

The agreement between the three annotators was measured using average pairwise percent agreement. The value of 57% was obtained for three-class annotation of the Hollow Earth data, whereas a value of 45% was obtained for the New World Order data. In the case of the Chemtrails dataset, a value of 43% was determined. To compensate for bad percent agreement, we opted for a majority vote to determine the class. To also include comments that did not yield an agreement (i.e., that were coded as a different class by all three annotators) in the analysis, we asked another well-trained annotator to label the remaining comments. This procedure ensured that all 8,000 comments were used when training the model. The final distribution of the classes with their frequency is shown in Table 2.

Table 2. Labeled datasets with sentiment score and their numbers of samples.

	Datasets		
Class (sentiment)	Hollow Earth	Chemtrails	New World Order
Contra-theory	1,389	2,105	1,215
Pro-theory	928	2,383	1,719
Other	5,683	3,512	5,066
Total	8,000	8,000	8,000

To address the challenge of unbalanced class distribution, we included a further analysis in addition to the main analysis with the entire dataset, in which over-represented classes were undersampled to ensure an equal distribution (Online Appendix C). This not only guarantees the integrity of the subsequent results, but also ensures the comparability of the predictions of the network analyses in the later course. However, it should be mentioned that we will always refer to the entire dataset of the analyses in the further course of the work.

3.3. Bidirectional Encoder Representations from Transformers (BERT)

In recent years, the field of NLP has changed rapidly. New deep learning approaches have significantly advanced the state of the art by achieving higher accuracy scores in various

applications. We decided to use the current state-of-the-art model for NLP tasks, BERT [69], which can be used for common NLP tasks such as text classification, translation, summarization, and question-answering, and which outperformed previous machine learning techniques.

BERT, which is based on multiple transformer networks [58], uses stacked attention layers and allows training on unsupervised tasks by pre-training on a large corpus. Transformer layers allow words to be represented better in relation to all other words using self-attention to better memorize long-term dependencies in sequences. Since BERT is bidirectional and therefore uses a BiLSTM network, all parameters are represented in a way that makes them comparable to each other, allowing a higher degree of expression of the word embeddings in the corpus. In contrast to word2vec [59] and GloVe [60], which use context-free and vocabulary-based approaches, BERT represents the input as subwords of individual words that can be derived from the entire context. One of the most important advantages of BERT is its generalizability, which means that BERT models can easily be fine-tuned for various NLP tasks, especially when less data is available to solve domain-specific tasks more effectively than with conventional methods. For the training of the language model based on domain-specific data, an extra domain-specific layer is trained on the top layer of BERT using the fine-tuning process.

3.3.1. Pre-processing

We performed individual pre-processing steps on our dataset to improve the data quality and, thus, the prediction performance. Data was pre-processed differently for BERT and the baseline methods, since their requirements differ considerably. All comments and replies were converted into lowercase and hyperlinks replaced with the term “url.” Furthermore, all models (BERT and baselines) were split into training (80%) and test data (20%). In general, for the baseline process, we tokenized the words. Subsequently, we converted a collection of comments to a matrix of token counts and used TF-IDF (Term

Frequency–Inverse Document Frequency) to achieve a detailed word representation of important terms. This procedure was implemented in pipelines, which are fixed series of workflows of several tasks.

Using a pre-trained BERT model, the data pre-processing needs to be adapted to the model. First, we shortened the comments for all datasets to the maximum sequence length. To this end, we concentrated on the median of all comments. Since our datasets contained individual comments with a large sequence length of comments, the arithmetic mean is not appropriate. Since the median sequence lengths of the different datasets are between 85 and 130, we set the maximum length of the sequences to 128. Our trained BERT model needed more comprehensive preprocessing steps in order to be able to process the data. Therefore, we tokenized the data using the tf-hub model, which simplifies pre-processing. For this process, the words are converted to lowercase characters and then tokenized by WordPiece tokenization [61]. Therefore, words are split into small subwords, e.g., “believing” into “believe” and “‘###ing,” which guarantees that a wider spectrum of out-of-vocabulary (OOV) words can be covered. After tokenization, the vocabulary is initialized where the most common combinations of existing words in the vocabulary are added iteratively; if words do not exist in the vocabulary, they are represented by individual characters: #H#o#l#l#o#w#E#a#r#t#h. Finally, special tokens are added at the beginning and at the end of the sentence, making it possible to find a better semantic connection between the sequences using the attention layer. For example, the token “[CLS]” marks the beginning of the sentence, while punctuation marks or the end of sentences are marked with “[SEP].”

3.3.2. Fine-tuning

For our analysis, we applied the official uncased model, which was pre-trained on Wikipedia (2.5B words) and the BookCorpus (800M words) and includes 12-layer, 768-hidden, 12-heads, and 110M parameters. This BERT model is able to predict YouTube comments and replies that contain content on conspiracy theories by classifying three classes

(pro-theory, contra-theory, other). Concerning the training of the models, we set a batch size of 32, due to the fact that our dataset is not large enough and the classes are distributed unequally. Furthermore, we decided to evaluate four different epochs (1,2,3,4) in order to have comparative values within the BERT models. We set the learning rate to 2×10^{-5} with a warm-up proportion of 10% to gradually increase the small learning rate. Since this is a sequence classification task, the label probabilities are computed with a standard softmax output layer.

We applied the machine learning models Logistic Regression (LR) and Support Vector Machine (SVM) with a linear kernel based on the LIBSVM implementation [62] to a TF-IDF weighted bag of words as baseline approaches. The hyperparameter search used a grid search with five-fold cross-validation to find the best parameters. Domain-specific models were built separately for each dataset.

3.3.3. Evaluation

Table 3 shows the results from the prediction of the test dataset to compare the applied models with each other using the weighted average and macro-average metric of the F1 score. The comparison illustrates that BERT achieves the best results within the three datasets and, thus, outperforms the baseline models. A detailed illustration of the prediction within each class of BERT can be found in Table 4. In particular, it can be seen that reaching a good accuracy is more challenging for the "contra-theory" class in the New World Order dataset. However, this problem does not seem to be due to the model, since the baseline models reveal the same patterns. For this reason, one can assume that this problem is due to the unbalanced dataset and its small number of trained records and that better results could be achieved with additional datasets. Based on our analysis using the undersampled dataset, we found that the range of F1 scores between the three classes decreased. However, the results also show that the value of the weighted average F1 score decreased overall due to the data reduction (see Online Appendix C). When these results are compared to the baseline models, it is noticeable

that the results of some baseline models still perform better than those of BERT models trained in only one epoch.

Table 3. Model evaluation of deep learning and machine learning methods on the test dataset.

			Macro			**Weighted**		
Dataset	**Models**	**Epoch**	**Precision**	**Recall**	**F1 Score**	**Precision**	**Recall**	**F1 Score**
Hollow Earth	**BERT**	1	0.570	0.460	0.465	0.694	0.741	0.695
		2	**0.592**	**0.562**	**0.575**	**0.727**	**0.743**	**0.734**
		3	0.585	0.562	0.571	0.721	0.736	0.727
		4	0.591	0.537	0.554	0.718	0.743	0.726
	LR	-	0.533	0.557	0.542	0.712	0.684	0.696
	SVM	-	0.533	0.554	0.541	0.714	0.689	0.700
Chemtrails	**BERT**	1	0. 596	0.575	0.570	0.605	0.617	0.597
		2	**0.595**	**0.583**	**0.583**	**0.606**	**0.617**	**0.606**
		3	0.574	0.571	0.571	0.588	0.594	0.590
		4	0.559	0.559	0.559	0.578	0.579	0.578
	LR	-	0.552	0.549	0.549	0.568	0.575	0.570
	SVM	-	0.559	0.558	0.558	0.575	0.578	0.576
New World Order	**BERT**	1	0.411	0.449	0.423	0.561	0.678	0.610
		2	0.541	0.501	0.507	0.636	0.671	0.645
		3	**0.539**	**0.507**	**0.517**	**0.634**	**0.660**	**0.643**
		4	0.531	0.508	0.514	0.633	0.656	0.641
	LR	-	0.523	0.503	0.508	0.627	0.652	0.636
	SVM	-	0.496	0.508	0.500	0.626	0.603	0.613

Table 4. Summary of the precision, recall, and F1 score for each class. Prediction based on the final BERT models (second epoch) to predict user-generated comments for each conspiracy theory.

		Metrics				
Dataset	Sentiment	Precision	Recall	F1 Score	Support	Prediction
Hollow Earth	Contra-theory	0.507	0.403	0.449	278	221
	Pro-theory	0.440	0.400	0.419	190	173
	Neither	0.830	0.884	0.856	1,132	1,206
	Weighted avg.	0.727	0.743	0.734	1,600	1,600
Chemtrails	Contra-theory	0.530	0.392	0.451	426	315
	Pro-theory	0.587	0.556	0.571	491	465
	Neither	0.667	0.801	0.728	683	820
	Weighted avg.	0.606	0.617	0.606	1,600	1,600
New World Order	Contra-theory	0.361	0.249	0.295	245	169
	Pro-theory	0.514	0.434	0.470	346	292
	Neither	0.742	0.837	0.787	1,009	1,139
	Weighted avg.	0.634	0.660	0.643	1,600	1,600

After the results of the BERT classifier on the test datasets were found to be good, the labels of all comments and replies to the conspiracy theories could be predicted. An overview can be found in Table 5. For the further course, we decided to use the BERT model with the two epochs for the datasets Hollow Earth and Chemtrails and with the third epoch for the New World Order dataset as a basis for further predictions. To take into account the fact that a user can write multiple comments, we considered the probabilities of each class for all written comments and calculated, for each user, the average probability of each class over their comments. Each user was assigned the class that was the most likely on average. This makes it possible to condense a user´s entire communication history into a single value, which is helpful for visualization purposes. This representation of the users is especially important for building the network, as well as for the calculation of opinion-based homogeneity.

Table 5. Predicted sentiment and numbers of comments of the whole dataset with the trained BERT models.

	Dataset		
Class (sentiment)	Hollow Earth	Chemtrails	New World Order
Contra-theory	4,900	19,437	11,999
Pro-theory	2,962	17,555	13,184
Other	7,450	25,239	20,916

3.4. Network analysis

To calculate the opinion-based homogeneity, we converted the data into a network as follows: Each YouTube video is a node, and each user who commented on at least one of the videos is also a node. Edges represent interactions, that is, two nodes are linked by a directed edge from node A to node B if user A has commented on video B or if user A has replied to a comment made by user B. In the resulting network, video nodes tend to be hubs, since videos typically receive many more comments than the typical comment receives replies.

We determined the stance of each node towards the respective conspiracy theory (pro, contra, or other). The stance of video nodes had already been determined by manual annotation (see Section 3.2.1). For user nodes, the classifier outputs for their individual comments were

aggregated using the arithmetic mean in order to take all their comments into account (compare [13]).

Nodes in the "other" class were removed since they were not relevant to studying the network relationships between supporters and opponents of the conspiracy theory. For the same reason, self-loops (comments on users' own videos and replies to their own comments) and isolated nodes (videos without comments and comments without replies by someone else) were removed.

We then calculated the global E-I index [54] and directed per-group E-I indices. In the calculation of per-class E-I indices, the direction of the edges was taken into account by only counting outgoing ties as external ties. As a result, the per-class index reflects the choices of the members of that group regarding who to interact with, and it therefore allows for a more accurate picture than the commonly used undirected group-wise E-I index.

To examine whether a given E-I index is significantly smaller or greater than would be expected if group members had no preference for internal or external ties, we used a permutation test (compare [63]). This sampling distribution of the E-I index is obtained by repeatedly rewiring each edge of the graph. This method keeps the number of nodes in each group constant, as well as the number of ties in the network (and thus its overall density). It thereby tests the null hypothesis that edges are distributed at random between the nodes.

To generate a network structure from the collected YouTube data and, thus, to calculate the opinion-based homogeneity, the data must first be converted. This mapping of the network allows a detailed inclusion of videos in the network that distribute a particular opinion, as well as users who respond to the video with comments. The relevant nodes, which are thus represented as hubs, are also included in the calculation of the E-I index. It is important that these hubs also have a stance, since they act as key players in the network and are likely to mark the general valence of the discussion. Users can also react to users to stimulate discussion and respond to different or similar opinions.

4. Results

The annotation of the videos on three conspiracy theories revealed that, in our YouTube dataset, the most common videos were those that supported the theory rather than debunk it. In relative numbers this means that videos on YouTube supporting conspiracy theories (58–81%) are more prevalent than videos that oppose such theories (8%–33%). In particular, a comparison of the three conspiracy theories shows that, in the "New World Order" dataset, videos that support the theory are clearly more prevalent (81%) than those counter-arguing the theory (7.94%). The conspiracy theories "Hollow Earth" and "Chemtrails" have a similar distribution of supporting (61.54%, 58.21%), debunking (29.23%, 32.84%), and *neither* videos (9.23%, 11.17%). Figure 1 shows the distribution of the categories in the videos, showing that in all theories, conspiracy-supportive content is more common than counter-conspiracy videos.

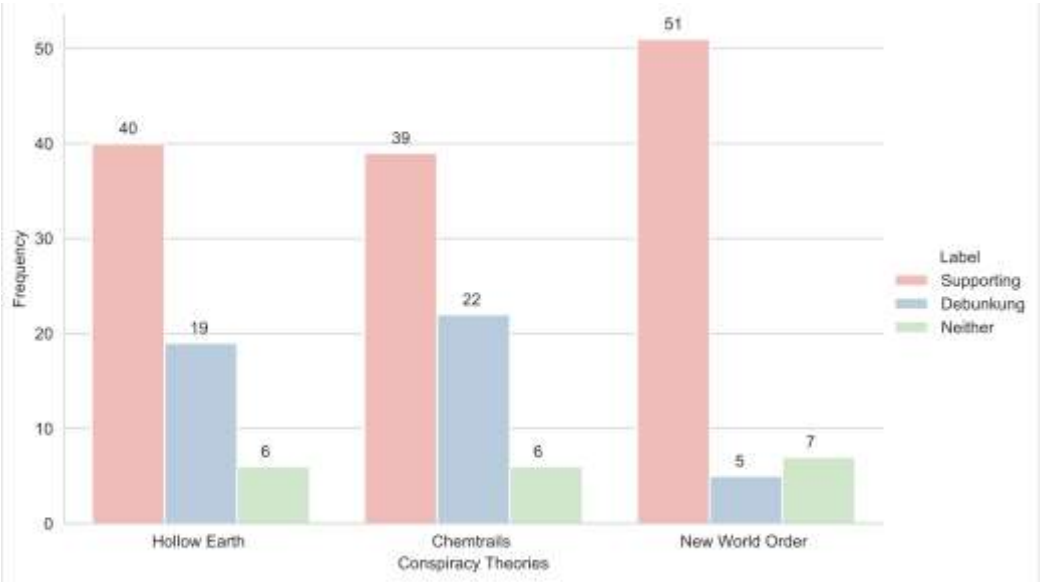

Figure 1. Graphical representation of the distribution of conspiracy videos with their stance. Note*: Since some videos were deleted from YouTube, the distribution of conspiracy videos is unequal.*

4.1. Popularity indicators of conspiracy theories

To address RQ1, altogether, N = 176 YouTube videos (130 conspiracy; 46 counter-conspiracy) are included in the analysis. Likes for the conspiracy group ranged from 0 to 224,881 (M = 3,533.35, SD = 20,091.25), dislikes from 0 to 5,649 (M = 231.25, SD =

677.07), views from 3 to 6,333,156 (M = 249,207.56, SD = 737,601.33), and comments from 0 to 50,705 (M = 1,032.36, SD = 4,672.59). For the counter-conspiracy group, the likes ranged from 0 to 34,300 (M = 2,076.8, SD = 5,246.45), dislikes from 0 to 3,704 (M = 416.57, SD = 749.47), views from 2 to 4,192,952 (M = 617,125.45, SD = 165,564.33), and comments from 0 to 8,898 (M = 1,146.3, SD = 1,978.62). Comparing the standard deviation in number of likes, dislikes, views, and comments shows that there exist huge differences within each group. We used an independent Mann-Whitney U test to compare popularity indicators such as likes, dislikes, comments, and views between conspiracy videos and counter-conspiracy videos. There was no significant difference in the numbers of: (a) Likes (W = 3,093, z = 0.61, p = 0.73), (b) dislikes (W = 2,607.5, z = -0.85, p = 0.197) and (c) comments (W = 2,617.5, z = -0.81, p = 0.209), but there was a significant difference in the number of (d) views (W = 3,639.5, z = -1.90, p = 0.029*). For the latter, means indicate that conspiracy videos are viewed significantly more frequently than counter-conspiracy videos. This effect, though, was small in magnitude (as specified by r). For three out of the four indicators, the significance test does not reject this null hypothesis. Therefore, although conspiracy-supportive videos are more prevalent, it seems that there is a balanced distribution of user reactions related to both conspiracy- and counter-conspiracy videos. The results of the test are shown in Table 6 (where *S.D.* is standard deviation, *W* is the Wilcoxon test statistic, *z* is the z-score, *p* is probability, and *r* is the effect size).

Table 6. Differences between conspiracy (N=130) and counter-conspiracy (N=46) videos. *Note: (C) represents conspiracy and (C-C) represents counter-conspiracy. *P<0.05.*

Measured	Group	Mean	S.D.	*W*	*z*	*p*	*r*
Likes	C	3,533.35	20,091.25	3,093	0.61	0.73	0.046
	C-C	2,076.8	5,246.45				
Dislikes	C	231.25	677.07	2,607.5	-0.85	0.197	-0.064
	C-C	416.57	749.47				
Comments	C	1,032.36	4,672.59	2,617.5	-0.81	0.209	-0.061
	C-C	1,146.3	1,978.62				
Views	C	249,207.56	737,601.33	3,639.5	-1.90	0.029*	-0.143
	C-C	165,564.33	617,125.45				

The distribution of the different popularity indicators (likes, dislikes, views, comments) of the three conspiracy theories is graphically summarized by the group's conspiracy and counter-conspiracy in a box-whisker plot in Figure 2. Due to the strong fluctuations of the popularity indicators between the three different conspiracy theories, we decided to scale the data points based on the symmetric logarithm. The plot illustrates that the distribution of popularity indicators differs between the counter-conspiracy and conspiracy videos within the conspiracy theories. It shows that on average, counter-conspiracy videos on Hollow Earth and Chemtrails generate less attention, as measured by all popularity indicators, than conspiracy theory videos. Furthermore, as can be seen in the low average and median popularity indicators, the Hollow Earth conspiracy theory is the one that has generated the least attention. In contrast, the New World Order conspiracy theory shows that, on average, the values of the popularity indicators are higher for the counter-conspiracy videos.

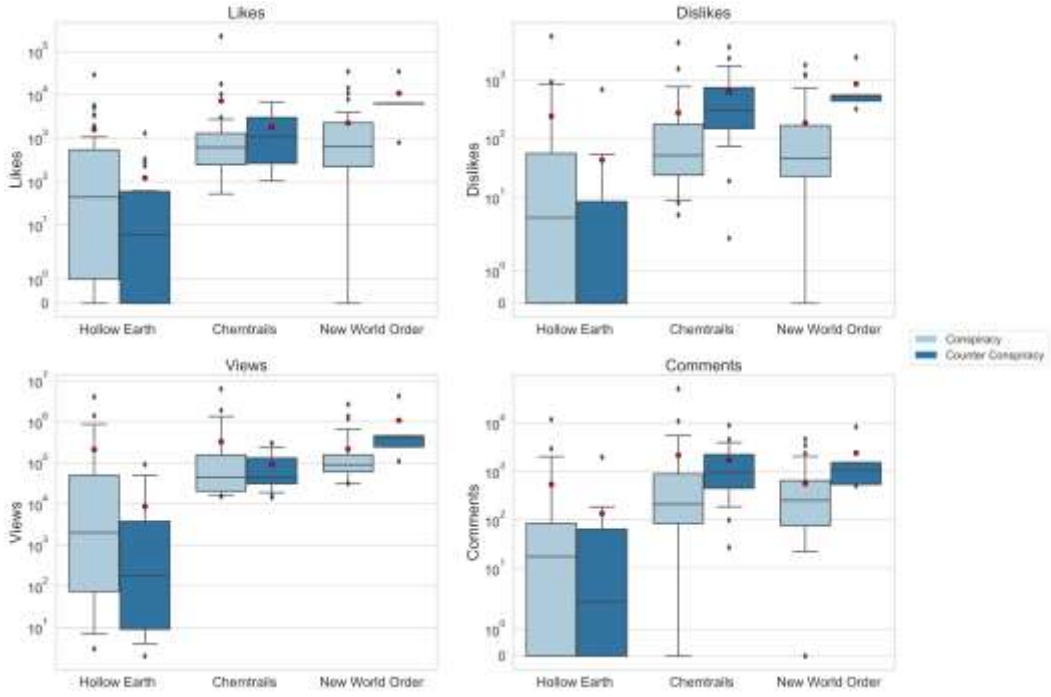

Figure 2. Clustered box-whisker-plot of popularity indicators on counter-conspiracy and conspiracy theories. The values of the popularity indicators (likes, dislikes, views, and comments) are displayed on a logarithmic axis. The black line indicates the median, the boxes the 25th and 75th percentiles and the whiskers extend to the 5th and 95th percentiles. The categories counter-conspiracy and conspiracy are indicated with color.

4.2. Prevalence of content including conspiracy theories

To examine RQ2, we used the 8,000 randomized and annotated user-generated comments to determine the distribution of the opinion climate (pro versus contra comments), which can be found in Table 2. This analysis shows that in two out of three cases, user-generated comments from supporters of the theory are more frequent than comments with a disapproving stance. However, this distribution was not found for the conspiracy theory of Hollow Earth, in which more comments included counter-messages than support of the theory.

4.3. Homogeneity and heterogeneity within discussion on conspiracy theories

Results related to opinion-based homogeneity among user-generated comments and videos on conspiracy theories (see RQ3), as the main interest of the present study, can be found in Tables 7 and 8. The results indicate that users who support the conspiracy theory are more likely to respond to videos and exchange comments with users that have the same opinion. This result can be shown by the class E-I index, which yielded negative values in all three datasets. Here, the datasets of the conspiracy theory Hollow Earth and New World Order are represented with the strongest negative E-I index values of –0.785 and –0.549, which shows relatively strong homogeneous interactions. The value of –0.221 in the dataset Chemtrails also represents a negative E-I index, but is not as strong as for the other two conspiracy theories.

Furthermore, it is noteworthy that for two out of three theories, people who advocated against the conspiracy theory show more heterogeneous communication behavior, except for the conspiracy theory Chemtrails. This is corroborated by the positive values of the E-I index: the dataset Hollow Earth has a value of 0.708, and New World Order of 0.377. The dataset of Chemtrails has a small positive value of 0.031. However, the value is close to 0 and can therefore be interpreted as neither homogeneous not heterogeneous.

These findings are mainly consistent with the results of the analyses of the undersampled dataset (see Online Appendix C) and have the same tendency, indicating that the unbalanced nature of the dataset does not have a significant influence on the prediction using BERT.

Considering the permutation test, the results in Table 8 further show that the expected E-I index is also negative for the "pro-theory" class and positive for the "contra-theory" class. The difference between the observed E-I index and the expected E-I index for the class "pro-theory" is 0.999 for the Hollow Earth dataset, 0.167 for Chemtrails, and 0.178 for New World Order. For the class "contra-theory," the difference between observed and expected E-I index for the Hollow Earth dataset is 0.919, for the Chemtrails dataset 0.357, and 0.005 for the New World Order dataset.

Regarding the results of the null hypothesis test, the values of the observed E-I index are significantly closer to –1 than expected for the pro-theory class in two out of the three datasets (Hollow Earth and New World Order), and significantly closer to +1 than expected in only one (Chemtrails). For the contra-theory class, they are significantly closer to –1 than expected for Chemtrails, but significantly closer to +1 than expected for Hollow Earth.

Table 7. Determining opinion-based homogeneity.

		Network statistics	
	Sentiment	Internal ties	External ties
Hollow Earth	Contra-theory	168	984
	Pro-theory	673	81
Chemtrails	Contra-theory	2,794	2,971
	Pro-theory	6,296	4,015
New World Order	Contra-theory	491	1,085
	Pro-theory	2,712	789

Table 8. Results of the permutation test with the observed and expected class E-I index. A permutation test with 1000 iterations was used to evaluate whether the observed index value was significantly higher ([P (obs >= exp)] or lower [P (obs <= exp)] than expected.

Sentiment	Observed E-I index	Expected E-I index	P (obs >= exp)	P (obs <= exp)

Hollow Earth	Global	0.118	-0.045	1.00	<0.01*
	Contra-theory	0.708	-0.211	1.00	<0.01*
	Pro-theory	-0.785	0.211	<0.01*	1.00
Chemtrails	Global	-0.131	-0.151	0.993	0.006*
	Contra-theory	0.031	0.388	<0.01*	1.00
	Pro-theory	-0.221	-0.388	1.00	<0.01*
New World Order	Global	-0.262	-0.137	<0.01*	1.00
	Contra-theory	0.377	0.372	0.582	0.396
	Pro-theory	-0.549	-0.371	<0.01*	1.00

For the graphical representation, the networks of the respective conspiracy theories are shown in Figures 3–5, where the nodes are marked with the classes (pro-theory, contra-theory) that represent an individual user or the published video on YouTube. We used Gephi [64] and the Force Atlas 2 layout algorithm [65] to visualize the networks. Nodes with the color green represent individuals who expressed support for the conspiracy theory, while red nodes represent advocates against the theory. Furthermore, we marked the edges starting from the source nodes with their color in order to highlight the communication paths. YouTube videos can be identified by their hub-like representation, indicating the highest in-degree. To generalize, we characterized videos that expressed support for the theory as pro-theory, while videos that disapproved of the theory were classified as contra-theory. The edges between the nodes reflect the lines of communication between the individual actors. We generated the network as a directed graph to see to whom the comments and answers are addressed. We have summarized the properties of the networks in Table 9 to provide a more comprehensive overview of the networks. The network properties show that the videos, comments, and replies related to the theory of Chemtrails make up the largest network. Similarly, the relatively high in-degree shows that there are very influential hubs in all networks. These hubs are the users who uploaded a video on a conspiracy theory and, thus, generated a lot of attention in the form of comments and replies. The out-degree indicates that the values of Hollow Earth and New World Order are very close to each other, while Chemtrails has a very high value of 412. The reason for this may be that an influential and highly active user has

commented on numerous videos or has replied to several comments from other users. Furthermore, it is noticeable that the conspiracy theories Hollow Earth and New World Order have a diameter (maximum distance between any pair of nodes) of two and three, while Chemtrails have a diameter of eight. This difference might be due to the size of the network and their different discussions. The density (i.e., the degree of interconnectedness) of our discussion networks shows a very small value in all three networks, which can be explained by the fact that the data are based on real networks and were transformed from videos and comments.

Table 9. Network properties.

	Datasets		
Network parameter	Hollow Earth	Chemtrails	New World Order
Nodes	1,864	11,484	5,000
Edges	1.906	16,076	5,077
Avg. degree	1.02	1.4	1.02
Diameter	2	8	3
Max. out-degree	20	412	8
Max. in-degree	374	1,099	497
Density	0.00055	0.00012	0.00020

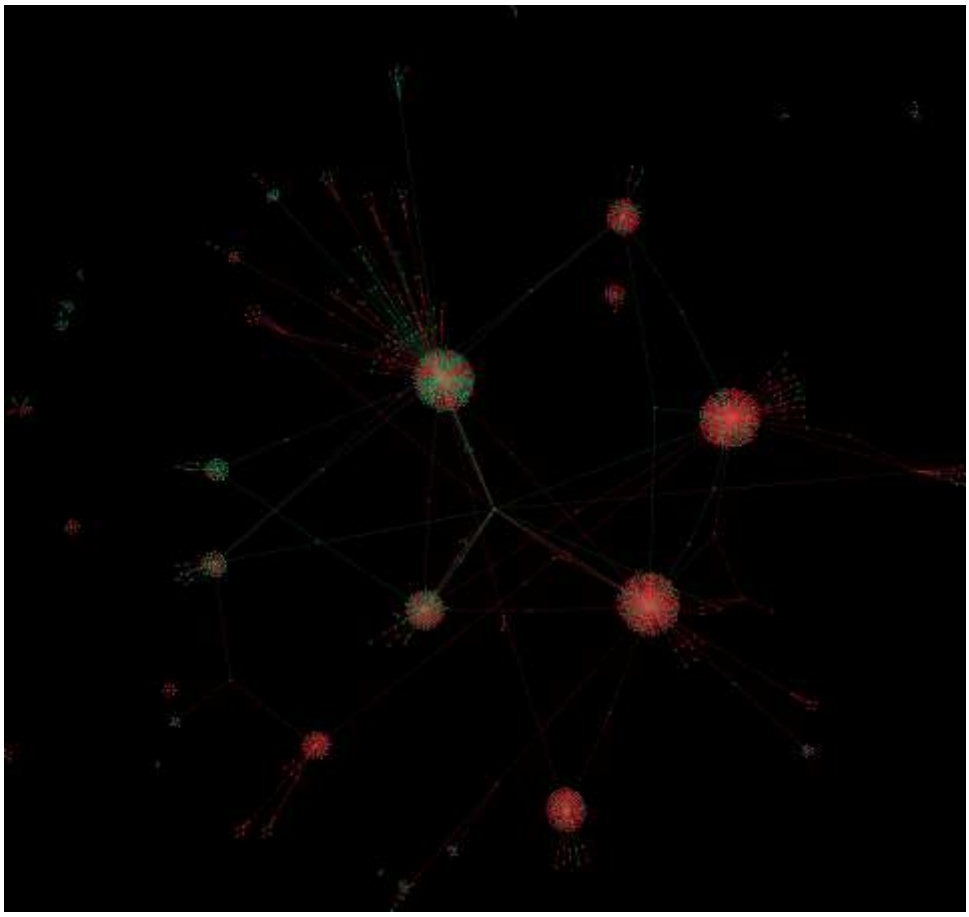

Figure 3. Discussion network of Hollow Earth. The network has 1,864 nodes, 1,906 edges and an average degree of 1.02. The visualization is based on the Force Atlas 2 layout algorithm. Green nodes represent individuals who expressed support for the hollow earth conspiracy theory; red nodes represent advocates against the hollow earth theory. The edges are colored according to the color source node.

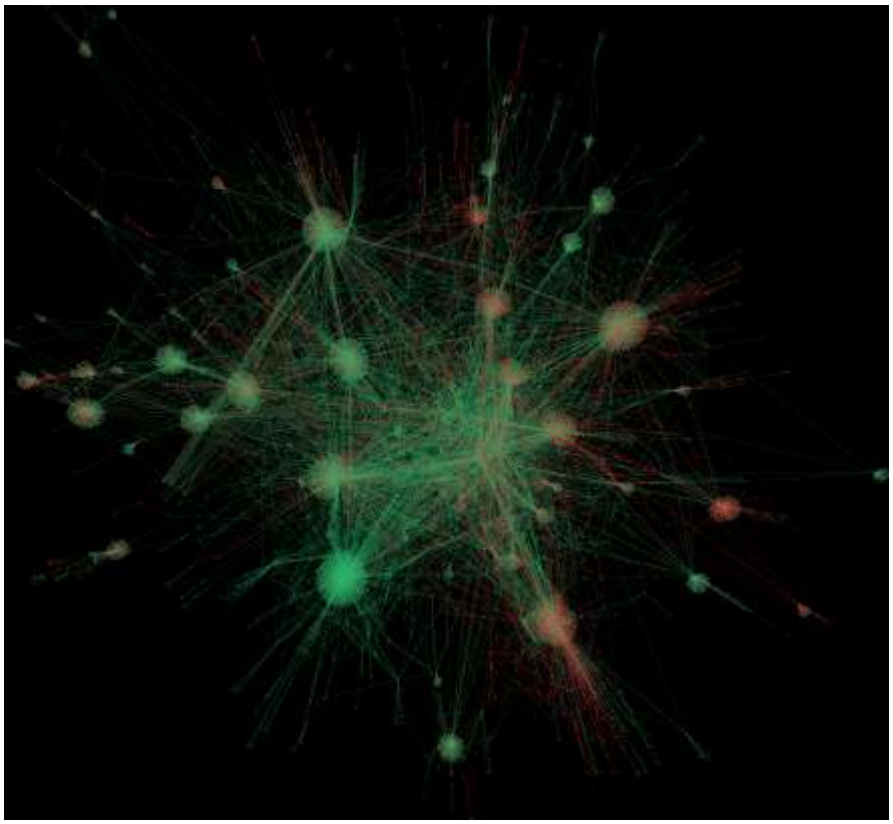

Figure 4. Discussion network of Chemtrails. The network has 11,484 nodes, 16,076 edges and an average degree of 1.4. The visualization is based on the Force Atlas 2 layout algorithm. Green nodes represent individuals who expressed support for the chemtrails conspiracy theory; red nodes represent advocates against the chemtrails theory. The edges are colored according to the color source node.

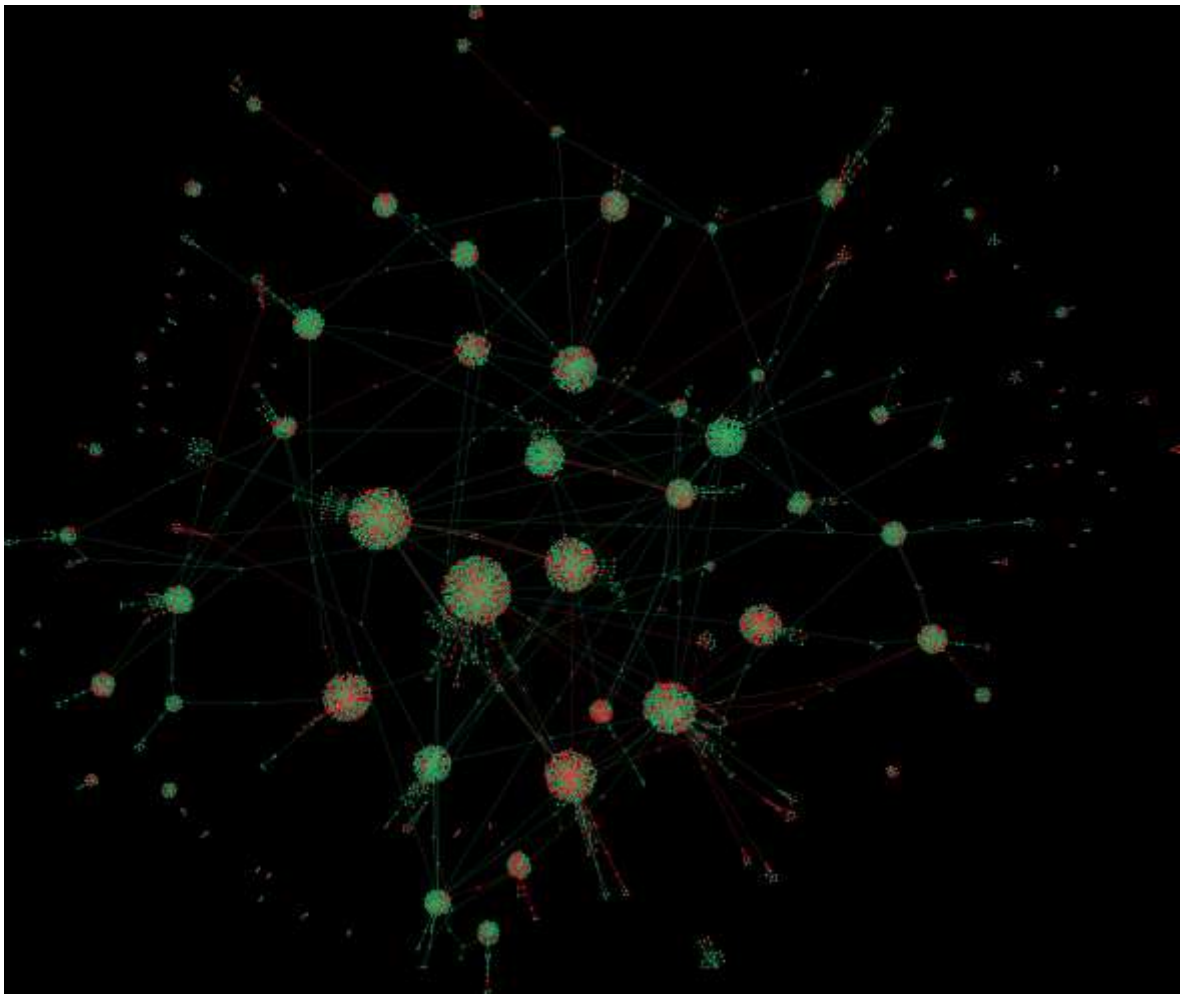

Figure 5. Discussion network of New World Order. The network has 5,000 nodes, 5,077 edges and an average degree of 1.02. The visualization is based on the Force Atlas 2 layout algorithm. Green nodes represent individuals who expressed support for the new world order conspiracy theory; red nodes represent advocates against the new world order theory. The edges are colored according to the color source node.

5. Discussion

The present work investigated the communication structure of conspiracy-related content in the form of videos and their user-generated comments on YouTube. Drawing on the spiral of silence theory and the status of a minority in a network, we were interested in the concept of opinion-based homogeneity in discussion networks related to conspiracy theories. As one of the first studies using this approach, the present work indicates that users on YouTube who express support for a conspiracy theory are more likely to engage in like-minded discussions than people who advocate against conspiracy theories.

5.1. User reactions on conspiracy and counter-conspiracy videos

With respect to RQ1, it was found that more user responses (likes, dislikes and comments) are given for conspiracy than for counter-conspiracy videos; however, these are comparatively small and statistically non-significant differences. Nevertheless, a statistically significant difference between the two groups was only found in the number of views. Previous studies evaluating conspiracy theories as misinformation observed that there is a risk that people will be indirectly influenced as a result of a high number of views, likes, or comments [66]. In this context, our findings give rise to optimism that, despite the dominance of conspiracy- over counter-conspiracy videos, the attention users pay (as measured by likes, dislikes, or comments) does not differ between the two types of videos. As mentioned in the literature review, Weeks & Gil de Zúñiga [35] suggest that countermeasures should be published by influential sources who also inspire trust and increase their social influence. Although only a small number of counter-conspiracy videos were found on YouTube, these influential videos could still contribute to the correction of conspiracy theories. Especially if these videos are characterized by a high number of popularity indicators such as likes, views, or comments, these videos could work against the narratives of conspiracy-supportive videos and reach a wide audience.

However, our results clearly showed that the number of videos featuring conspiracy-supportive content on YouTube, and thus actively contributing to the process of misinformation diffusion, is greater than the number of videos that debunked these conspiracy theories with facts. The increasing number of videos on conspiracy theories is alarming in view of the social problem of so-called filter bubbles, that is, the idea that (recommendation) algorithms shape the information landscape of users based on previous information selection patterns [34]. Thus, once users are exposed to the first video on a certain conspiracy theory, they can easily get recommendations about further videos on this theory [29]. This, in turn, could capture them in homogeneous information cocoons, reinforcing their conspiracy beliefs

and contributing to collective polarization [67]. Therefore, given the prevalence of conspiracy videos on YouTube, more research is needed to disentangle the recommendation patterns on platforms such as YouTube in order to estimate the probability that users who were either already interested in conspiracy beliefs, were incidentally exposed to the videos, or who were actually interested in viewing theory-debunking videos will encounter even more misinformation.

5.2. Minority opinions and their homogeneous discussion network

With regard to RQ2, we found that concerning the conspiracy theories Chemtrails and New World Order, there are more comments supporting than refuting these theories. Only the dataset on Hollow Earth contains more comments against the theory than in favor of the theory. Again, this distribution provides optimistic insights into the tone of discussion on this particular topic, suggesting that promoters of this theory are met with resistance in the form of commenters who advocate against the validity of this theory. Nevertheless, it remains unclear: (a) Whether theory supporters indeed encounter and read these comments, and (b) whether counter-comments include the characteristics that are necessary to successfully outline the falsehood of this theory [36], [68]. One needs to bear in mind that for two out of three cases, theory-supportive comments were more prevalent than contra-theory comments. An explanation for this could be that predominantly those who believe in the conspiracy feel the urge to discuss this theory on platforms such as YouTube. This, in turn, leads to an over-representation of support posted below conspiracy-related videos that may not represent the actual distribution of opinions among the population. Considering exemplification effects [41], this could lead to false inferences about “what most others may think” about this theory. In other words, if I see that many comments speak in favor of this theory described in a YouTube video, this could lead me to the conclusion that there is wide support in society for this conspiracy theory. This inaccurate inference could also shape my personal judgment, affecting—in the long run—my own belief in the theory [42]. Given that previous studies

indicated comparatively small, albeit significant, effects of comments on public opinion perceptions, one could assume that, in the context of conspiracy theories as niche topics, effects for regular citizens are even smaller. These speculations about potential effects of encountering user-generated comments supporting conspiracy beliefs need to be addressed systematically by future research (potentially by experimental studies) to disentangle which characteristics of the comments and their readers facilitate (the perception of) public acceptance and spread of misinformation in the form of conspiracy theories. Future work may also involve using the previous data to generate simulation models that grasp the dynamics within opinion climates on different conspiracy theories.

Regarding RQ3, and consistent with the spiral of silence theory [14], users who might perceive themselves as the minority in society prefer like-minded interactions over discussions on or responses to content or comments promoting a diverging stance on the conspiracy theory. Thus, those who expressed themselves in favor of the conspiracy theories interacted in more homogeneous networks. At the same time, this finding indicates that supposed supporters of conspiracy theories do not exhibit behavior that lets them, for instance, contradict the mainstream in discussions [15].

In contrast, users who challenge conspiracy theories interact in more heterogeneous discussion networks covering diverse opinions and, obviously, actively seek debate. The only exception were supposed opponents of the Chemtrails theory who seem to interact in neither homogeneous nor heterogeneous opinion networks. This pattern regarding the prevalence of opinion-based homogeneity extends prior research, which largely focused on political and controversial topics [13]. For three political issues, Röchert et al. [13] found evidence for relatively heterogeneous interactions among supporters and opponents within a political debate. The discrepancy between findings might be due to the nature of the topics: Conspiracy theorists (as analyzed in the present study) are often marginalized groups [9], [10] that might experience social rejection. As a special group in society, those who support these theories

might feel comfortable in homogeneous, more cohesive surroundings [2]. The homogeneity among supporters of conspiracy theories has implications for the discussion about potential fragmentation of online networks [11], [12]: In the long run, this potential segregation from heterogeneous interactions with challenging views could lead people with conspiracist worldviews to overestimate public support for a particular theory, feeling reinforced in their thinking. Whether this reinforcement leads to individual or collective polarization has yet to be examined by longitudinal approaches. Our findings, at least, indicate an asymmetry in the diversity of communication between those who support and those who oppose conspiracy theories. Why this pattern may vary across conspiracy theories requires further investigation by focusing on the specifics of each theory and their associated communities.

5.3. Limitations

One of the first limitations to be mentioned is the fact that only the 100 videos with the most views were crawled, which means that our data does not cover the entire discussion landscape of these topics on YouTube. In addition, the term "conspiracy" was used, which can be problematic unless people see their own theory as a conspiracy [25]. Leaving out this term would certainly have yielded more results, also covering niche networks on these conspiracies, but would also have led to many off-topic videos (e.g., "New World Order").

Due to the new YouTube guidelines in force, some videos collected in the previous step were later deleted in the course of the study. For this reason, we decided to exclude these data points from our analysis, since there are no longer any references to the original video material. Furthermore, the study shows a static snapshot of the current YouTube landscape of conspiracy-related content with its communication network, focusing on how supporters or opponents talk about these videos. Since these are not the only three conspiracy theories on YouTube, it seems worthwhile to see whether our findings on opinion-based homogeneity can also be replicated in the context of more controversial conspiracy theories.

A further limitation of the study is the partially unbalanced dataset, which makes the prediction of some classes (pro-theory, contra-theory) more difficult than the prediction of the class “others” (neither pro nor contra), and this is also reflected in the results. It should be noted, however, that this is a representation of reality, with the majority of users writing off-topic comments. Due to the state-of-the-art language model BERT, which we used for text classification, we were able to increase the prediction accuracy and, thus, also the generalizability of our models. However, one needs to bear in mind that comments can also be predicted incorrectly. This claim is based on the primary limitation that the multiclass classification of the individual classes (especially "Contra-theory" and "Pro-theory") have a relatively low F1 score. Reasons for these low F1 scores might be, on the one hand, that not enough training data was given or, on the other, that there are similar linguistic class features. For this reason, it is not advisable to propose this particular dataset as a standard benchmark dataset. It is important to note, nonetheless, that even in the intelligent human process of labeling data, disagreements occurred. This means, firstly, that it is apparently difficult even for humans to assign comments to an unambiguous stance, and secondly, that it is even more difficult for computers to predict these manually classified comments if human coders cannot get it right.

6. Conclusion

The present study has been one of the first attempts to thoroughly measure opinion-based homogeneity in the context of three conspiracy theories on YouTube. To this end, we combined a text classification approach with BERT and a network analysis to compute the E-I index to measure the homogeneity and heterogeneity of the network based on user comments. This study showed that for three conspiracy theories users who express support for those conspiracy beliefs are more likely to interact in homogeneous networks than are users who oppose those beliefs. This pattern found within discussion networks on YouTube offers new

insights for the debate on the fragmentation of social groups in online communication by specifying the (topical) circumstances under which homogeneity and potential segregation are likely to emerge. These findings have practical implications for platform providers who—by employing this methodical combination—could: (a) Detect particular sub-networks in which conspiracy beliefs are discussed and spread without any contradiction or correction, and (b) disseminate fact-checking messages in those sub-communities to counteract this ostensible legitimization of misinformation. Such an approach could help to strategically reach groups susceptible to believing in conspiracy theories in order to prevent them from becoming caught in homogeneous bubbles.

Declaration of Conflicting Interests

The authors declare that they have no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Software Information

We used a Python code (Version 3.6.7) for the analysis of the YouTube comments in order to sample our datasets and train BERT models; specifically, we used the Python packages: pandas, numpy, sqlalchemy, json, requests, pickle, re, string, datetime, random, scipy, nltk, matplotlib, seaborn, itertools, scipy, sklearn, keras, tensorflow, tensorflow-hub.
For the process of network analysis and the calculation of opinion-based homogeneity, we used an R-Script (Version 3.6.2) with the following packages: igraph, tidyverse, dplyr, xlsx, compute.es, esc, pbapply.

Data Availability

The developer policy of the YouTube Data API does not permit the publication or distribution of the data used in this study. To promote the replicability of our findings, we have provided a

detailed description of how to obtain the same or similar data in Online Appendix D. Please note that videos and user-generated comments used for the analyses in this study could be removed and future replications may not reach exactly the same results.

Supplemental Material

Supplemental material for this article is available online:

https://osf.io/adu6v/?view_only=a8332f25578c428cbe333e40e634e938

Funding

This research was supported by the Digital Society research program funded by the Ministry of Culture and Science of the German State of North Rhine-Westphalia (Grant Number: 005-1709-0004), Junior Research Group "Digital Citizenship in Network Technologies" (Project Number: 1706dgn009).

References

[1] A. Bessi, M. Coletto, G. A. Davidescu, A. Scala, G. Caldarelli, und W. Quattrociocchi, „Science vs conspiracy: Collective narratives in the age of misinformation", *PloS one*, Bd. 10, Nr. 2, S. e0118093, 2015.

[2] M. Del Vicario *u. a.*, „The spreading of misinformation online", *Proceedings of the National Academy of Sciences*, Bd. 113, Nr. 3, S. 554–559, 2016, doi: 10.1073/pnas.1517441113.

[3] J. E. Oliver und T. J. Wood, „Conspiracy theories and the paranoid style (s) of mass opinion", *American Journal of Political Science*, Bd. 58, Nr. 4, S. 952–966, 2014.

[4] K. M. Douglas, R. M. Sutton, D. Jolley, und M. J. Wood, „The social, political, environmental, and health-related consequences of conspiracy theories", *The psychology of conspiracy*, S. 183–200, 2015.

[5] V. Swami und A. Furnham, „12 Political paranoia and conspiracy theories", *Power, politics, and paranoia: Why people are suspicious of their leaders*, S. 218, 2014.

[6] D. Jolley und K. M. Douglas, „The social consequences of conspiracism: Exposure to conspiracy theories decreases intentions to engage in politics and to reduce one's carbon footprint", *British Journal of Psychology*, Bd. 105, Nr. 1, S. 35–56, 2014.

[7] S. van der Linden, „The conspiracy-effect: Exposure to conspiracy theories (about global warming) decreases pro-social behavior and science acceptance", *Personality and Individual Differences*, Bd. 87, S. 171–173, 2015.

[8] D. Jolley und K. M. Douglas, „The effects of anti-vaccine conspiracy theories on vaccination intentions", *PloS one*, Bd. 9, Nr. 2, S. e89177, 2014.

[9] J.-W. van Prooijen, J. Staman, und A. P. Krouwel, „Increased conspiracy beliefs among ethnic and Muslim minorities“, *Applied cognitive psychology*, Bd. 32, Nr. 5, S. 661–667, 2018.
[10] M. J. Wood und K. M. Douglas, „Online communication as a window to conspiracist worldviews“, *Frontiers in psychology*, Bd. 6, S. 836, 2015.
[11] A. Bruns, „It's not the technology, stupid: How the `Echo Chamber' and `Filter Bubble' metaphors have failed us“, 2019.
[12] C. R. Sunstein, *#Republic: divided democracy in the age of social media*. Princeton ; Oxford: Princeton University Press, 2017.
[13] D. Röchert, G. Neubaum, B. Ross, F. Brachten, und S. Stieglitz, „Opinion-based Homogeneity on YouTube : Combining Sentiment and Social Network Analysis“, *Computational Communication Research*, Bd. 2, Nr. 1, S. 81–108, Feb. 2020, doi: 10.5117/CCR2020.1.004.ROCH.
[14] E. Noelle-Neumann, „The Spiral of Silence A Theory of Public Opinion“, *Journal of Communication*, Bd. 24, Nr. 2, S. 43–51, 1974, doi: 10.1111/j.1460-2466.1974.tb00367.x.
[15] R. Imhoff und P. K. Lamberty, „Too special to be duped: Need for uniqueness motivates conspiracy beliefs“, *European Journal of Social Psychology*, Bd. 47, Nr. 6, S. 724–734, 2017.
[16] M. Cinelli *u. a.*, „The COVID-19 social media infodemic“, *Sci Rep*, Bd. 10, Nr. 1, S. 16598, Dez. 2020, doi: 10.1038/s41598-020-73510-5.
[17] J. Shin, L. Jian, K. Driscoll, und F. Bar, „The diffusion of misinformation on social media: Temporal pattern, message, and source“, *Computers in Human Behavior*, Bd. 83, 2018, doi: 10.1016/j.chb.2018.02.008.
[18] M. J. Wood, „Propagating and Debunking Conspiracy Theories on Twitter During the 2015–2016 Zika Virus Outbreak“, *Cyberpsychology, Behavior, and Social Networking*, Bd. 21, Nr. 8, S. 485–490, 2018, doi: 10.1089/cyber.2017.0669.
[19] A. Bessi, „On the statistical properties of viral misinformation in online social media“, *Physica A: Statistical Mechanics and its Applications*, Bd. 469, S. 459–470, 2017, doi: https://doi.org/10.1016/j.physa.2016.11.012.
[20] G. Donzelli *u. a.*, „Misinformation on vaccination: A quantitative analysis of YouTube videos“, *Human Vaccines & Immunotherapeutics*, Bd. 14, Nr. 7, S. 1654–1659, 2018, doi: 10.1080/21645515.2018.1454572.
[21] B. Monsted und S. Lehmann, *Algorithmic Detection and Analysis of Vaccine-Denialist Sentiment Clusters in Social Networks*. 2019.
[22] W. Kim, O.-R. Jeong, und S.-W. Lee, „On social Web sites“, *Information Systems*, Bd. 35, Nr. 2, S. 215–236, Apr. 2010, doi: 10.1016/j.is.2009.08.003.
[23] N. Newman, R. Fletcher, A. Kalogeropoulos, und R. Nielsen, *Reuters institute digital news report 2019*, Bd. 2019. Reuters Institute for the Study of Journalism, 2019.
[24] Pew Research Center, *For Local News, Americans Embrace Digital but Still Want Strong Community Connection*. 2019.
[25] B. L. Keeley, „Of Conspiracy Theories“, *The Journal of Philosophy*, Bd. 96, Nr. 3, S. 109–126, 1999.
[26] M. Wood und K. Douglas, „“What about building 7?” A social psychological study of online discussion of 9/11 conspiracy theories“, *Frontiers in Psychology*, Bd. 4, S. 409, 2013.
[27] U. Ahmad, A. Zahid, M. Shoaib, und A. AlAmri, „HarVis: An integrated social media content analysis framework for YouTube platform“, *Information Systems*, Bd. 69, S. 25–39, Sep. 2017, doi: 10.1016/j.is.2016.10.004.
[28] A. Bessi *u. a.*, „Users polarization on Facebook and Youtube“, *PloS one*, Bd. 11, Nr. 8, S. e0159641, 2016.

[29] J. Allgaier, „Science on YouTube: What user find when they search for climate science and climate manipulation“, *CoRR*, Bd. abs/1602.02692, 2016.
[30] A. Nerghes, P. Kerkhof, und I. Hellsten, „Early Public Responses to the Zika-Virus on YouTube: Prevalence of and Differences Between Conspiracy Theory and Informational Videos“, in *Proceedings of the 10th ACM Conference on Web Science*, 2018, S. 127–134.
[31] T. Goertzel, „Belief in conspiracy theories“, *Political Psychology*, S. 731–742, 1994.
[32] V. Swami *u. a.*, „Conspiracist ideation in Britain and Austria: Evidence of a monological belief system and associations between individual psychological differences and real-world and fictitious conspiracy theories“, *British Journal of Psychology*, Bd. 102, Nr. 3, S. 443–463, 2011, doi: 10.1111/j.2044-8295.2010.02004.x.
[33] S. Lewandowsky, K. Oberauer, und G. E. Gignac, „NASA Faked the Moon Landing—Therefore, (Climate) Science Is a Hoax: An Anatomy of the Motivated Rejection of Science“, *Psychological Science*, Bd. 24, Nr. 5, S. 622–633, 2013, doi: 10.1177/0956797612457686.
[34] J. B. Schmitt, D. Rieger, O. Rutkowski, und J. Ernst, „Counter-messages as Prevention or Promotion of Extremism?! The Potential Role of YouTube“, *Journal of Communication*, Bd. 68, Nr. 4, S. 780–808, Aug. 2018, doi: 10.1093/joc/jqy029.
[35] B. E. Weeks und H. Gil de Zúñiga, „What's Next? Six Observations for the Future of Political Misinformation Research“, *American Behavioral Scientist*, S. 0002764219878236, 2019, doi: 10.1177/0002764219878236.
[36] L. Bode und E. K. Vraga, „In Related News, That Was Wrong: The Correction of Misinformation Through Related Stories Functionality in Social Media“, *Journal of Communication*, Bd. 65, Nr. 4, S. 619–638, 2015, doi: 10.1111/jcom.12166.
[37] D. Jolley und K. M. Douglas, „Prevention is better than cure: Addressing anti-vaccine conspiracy theories“, *Journal of Applied Social Psychology*, Bd. 47, Nr. 8, S. 459–469, 2017, doi: 10.1111/jasp.12453.
[38] J. B. Walther und J. Jang, „Communication processes in participatory websites“, *Journal of Computer-Mediated Communication*, Bd. 18, Nr. 1, S. 2–15, 2012.
[39] S. Chaiken, „The heuristic model of persuasion“, in *Social influence: the ontario symposium*, 1987, Bd. 5, S. 3–39.
[40] S. S. Sundar, A. Oeldorf-Hirsch, und Q. Xu, „The bandwagon effect of collaborative filtering technology“, in *CHI'08 extended abstracts on Human factors in computing systems*, 2008, S. 3453–3458.
[41] D. Zillmann und H.-B. Brosius, *Exemplification in communication: The influence of case reports on the perception of issues*. Routledge, 2012.
[42] G. Neubaum und N. C. Krämer, „Monitoring the Opinion of the Crowd: Psychological Mechanisms Underlying Public Opinion Perceptions on Social Media“, *Media Psychology*, Bd. 20, Nr. 3, S. 502–531, Juli 2017, doi: 10.1080/15213269.2016.1211539.
[43] E.-J. Lee und Y. J. Jang, „What do others' reactions to news on Internet portal sites tell us? Effects of presentation format and readers' need for cognition on reality perception“, *Communication research*, Bd. 37, Nr. 6, S. 825–846, 2010.
[44] J. B. Walther, G. Neubaum, L. Rösner, S. Winter, und N. C. Krämer, „The effect of bilingual congruence on the persuasive influence of videos and comments on YouTube“, *Journal of Language and Social Psychology*, Bd. 37, Nr. 3, S. 310–329, 2018.
[45] S. Winter, „Impression-motivated News Consumption–Are user comments in social media more influential than on news sites“, *Journal Of Media Psychology*, S. 1864–1105, 2018.
[46] E. Bakshy, S. Messing, und L. A. Adamic, „Exposure to ideologically diverse news and opinion on Facebook“, *Science*, Bd. 348, Nr. 6239, S. 1130–1132, 2015, doi: 10.1126/science.aaa1160.

[47] A. Boutyline und R. Willer, „The social structure of political echo chambers: Variation in ideological homophily in online networks“, *Political Psychology*, Bd. 38, Nr. 3, S. 551–569, 2017, doi: 10.1111/pops.12337.
[48] E. Colleoni, A. Rozza, und A. Arvidsson, „Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data“, *Journal of Communication*, Bd. 64, Nr. 2, S. 317–332, 2014, doi: 10.1111/jcom.12084.
[49] E. Dubois und G. Blank, „The echo chamber is overstated: the moderating effect of political interest and diverse media“, *Information, Communication & Society*, Bd. 21, Nr. 5, S. 729–745, 2018.
[50] C. Vaccari, A. Valeriani, P. Barberá, J. T. Jost, J. Nagler, und J. A. Tucker, „Of echo chambers and contrarian clubs: Exposure to political disagreement among German and Italian users of Twitter“, *Social Media+ Society*, Bd. 2, Nr. 3, S. 2056305116664221, 2016.
[51] B. R. Warner und R. Neville-Shepard, „Echoes of a conspiracy: Birthers, truthers, and the cultivation of extremism“, *Communication Quarterly*, Bd. 62, Nr. 1, S. 1–17, 2014.
[52] S. Moscovici, E. Lage, und M. Naffrechoux, „Influence of a Consistent Minority on the Responses of a Majority in a Color Perception Task“, *Sociometry*, Bd. 32, Nr. 4, S. 365–380, 1969.
[53] W. Wood, S. Lundgren, J. A. Ouellette, S. Busceme, und T. Blackstone, „Minority influence: a meta-analytic review of social influence processes.“, *Psychological bulletin*, Bd. 115, Nr. 3, S. 323, 1994.
[54] D. Krackhardt und R. N. Stern, „Informal Networks and Organizational Crises: An Experimental Simulation“, *Social Psychology Quarterly*, Bd. 51, Nr. 2, S. 123–140, 1988, doi: 10.2307/2786835.
[55] A. Spark, „Conjuring order: the new world order and conspiracy theories of globalization“, *The Sociological Review*, Bd. 48, Nr. 2_suppl, S. 46–62, 2000.
[56] D. Stupple und A. Dashti, „Flying Saucers and Multiple Realities: A Case Study in Phenomenological Theory“, *The Journal of Popular Culture*, Bd. 11, Nr. 2, S. 479–493, 1977, doi: 10.1111/j.0022-3840.1977.00479.x.
[57] A. F. Wilson, „The bitter end: apocalypse and conspiracy in white nationalist responses to the Islamic State attacks in Paris“, *Patterns of Prejudice*, Bd. 51, Nr. 5, S. 412–431, 2017, doi: 10.1080/0031322X.2017.1398963.
[58] A. Vaswani *u. a.*, „Attention is all you need“, in *Advances in neural information processing systems*, 2017, S. 5998–6008.
[59] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, und J. Dean, *Distributed Representations of Words and Phrases and their Compositionality*. 2013.
[60] J. Pennington, R. Socher, und C. D. Manning, „GloVe: Global Vectors for Word Representation“, in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014, S. 1532–1543.
[61] M. Johnson *u. a.*, „Google’s Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation“, *Transactions of the Association for Computational Linguistics*, Bd. 5, S. 339–351, 2017, doi: 10.1162/tacl_a_00065.
[62] C.-C. Chang und C.-J. Lin, „LIBSVM: A Library for Support Vector Machines“, *ACM Transactions on Intelligent Systems and Technology (TIST)*, Bd. 2, Nr. 3, S. 27, 2011, doi: 10.1145/1961189.1961199.
[63] J. Scott und P. J. Carrington, Hrsg., *The SAGE handbook of social network analysis*. London ; Thousand Oaks, Calif: SAGE, 2011.
[64] M. Bastian, S. Heymann, M. Jacomy, und others, „Gephi: an open source software for exploring and manipulating networks.“, *Icwsm*, Bd. 8, Nr. 2009, S. 361–362, 2009.

[65] M. Jacomy, T. Venturini, S. Heymann, und M. Bastian, „ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software“, *PloS one*, Bd. 9, Nr. 6, S. e98679, 2014.
[66] K. M. Douglas und R. M. Sutton, „The Hidden Impact of Conspiracy Theories: Perceived and Actual Influence of Theories Surrounding the Death of Princess Diana“, *The Journal of Social Psychology*, Bd. 148, Nr. 2, S. 210–222, 2008, doi: 10.3200/SOCP.148.2.210-222.
[67] C. R. Sunstein, „Is Social Media Good Or Bad for Democracy“, *SUR-Int'l J. on Hum Rts.*, Bd. 27, S. 83, 2018.
[68] G. Orosz, P. Krekó, B. Paskuj, I. Tóth-Király, B. B\Hothe, und C. Roland-Lévy, „Changing Conspiracy Beliefs through Rationality and Ridiculing“, *Frontiers in Psychology*, Bd. 7, S. 1525, 2016, doi: 10.3389/fpsyg.2016.01525.
[69] Devlin, J., CHang, M. W., Lee, K. & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.