Proceedings of the Twentieth International AAAI Conference on Web and Social Media (ICWSM 2026)

Anti-Establishment Sentiment on TikTok: Implications for Understanding Influence(rs) and Expertise on Social Media

Tianliang Xu[1], **Ariel Hasell**[2], **Sabina Tomkins**[1]

[1] University of Michigan School of Information
[2]University of Michigan Department of Communication and Media
tianlix@umich.edu, hasell@umich.edu, stomkins@umich.edu

Abstract

Distrust of public serving institutions and anti-establishment views are on the rise (especially in the U.S.). As people turn to social media for information, it is imperative to understand whether and how social media environments may be contributing to distrust of institutions. In social media, content creators, influencers, and other opinion leaders often position themselves as having expertise and authority on a range of topics from health to politics, and in many cases devalue and dismiss institutional expertise to build a following and increase their own visibility. However, the extent to which this content appears and whether such content increases engagement is unclear. This study analyzes the prevalence of anti-establishment sentiment (AES) on the social media platform TikTok. Despite its popularity as a source of information, TikTok remains relatively understudied and may provide important insights into how people form attitudes towards institutions. We employ a computational approach to label TikTok posts as containing AES or not across topical domains where content creators tend to frame themselves as experts: finance and wellness. As a comparison, we also consider the topic of conspiracy theories, where AES is expected to be common. We find that AES is most prevalent in conspiracy theory content, and relatively rare in content related to the other two topics. However, we find that engagement patterns with such content vary by area, and that there may be platform incentives for users to post content that expresses anti-establishment sentiment.

Code — https://github.com/politechlab/AES

Introduction

In recent years, growing support for populism and anti-intellectualism have drawn increased scholarly attention to the role anti-establishment views play in undermining the health of democratic societies (Droste 2021; Merkley 2020; Oliver and Rahn 2016).

In the U.S., citizens across the political spectrum increasingly distrust public serving institutions like government, the news media, scientists, and universities. Many have developed views that are hostile towards those institutions, with consequences ranging from increased misinformation to political violence (Armaly and Enders 2024; Enders et al. 2023; Enders and Uscinski 2021; Lyons 2023; Stecula, Kuru, and Jamieson 2020). To understand and address large scale, global problems like public health crises, climate change, and democratic backsliding, there is a need to better understand the factors that might influence public opinion regarding public serving institutions and democratic systems.

Anti-establishment views include an array of attitudes including populism, anti-intellectualism, and conspiratorial thinking, among others (Enders and Uscinski 2021; Oliver and Rahn 2016; Uscinski et al. 2021), and can broadly be defined as "the politics of opposition to those wielding power"(Barr 2009, page 31). Such views position the "virtuous" ordinary citizens against the "corrupt" and "immoral" elite establishment, and generally the establishment can refer to any cultural, political, or economic elite (Droste 2021; Mudde 2004).

Importantly, the perception of who or what is part of the elite establishment can vary over time or across individuals, but these views consistently position everyday or ordinary people as morally better than experts or those working within elite institutions. For example, while trust in science is generally high in the U.S., there is increasing sentiment that scientists and universities are a part of a powerful, elite group and are attempting to be arbiters of "truth" while dismissing and undermining "common sense", in an effort to take power away from ordinary people (Mede and Schäfer 2020). Anti-establishment views can correlate with political ideology in certain contexts, but research shows that anti-establishment views function orthogonally to the traditional liberal-conservative political spectrum in the U.S. (e.g., Uscinski et al. 2021). The extent to which Americans hold anti-establishment views has varied over time, and while the majority of citizens do not hold such views, the prevalence of these views is increasing (Droste 2021; Mede and Schäfer 2020; Merkley 2020).

Most research in this area has focused on anti-establishment views aimed at the political and economic elite, like populism (e.g., Uscinski et al. 2021), but there has been increased attention towards views that are hostile towards other perceived elites, like scientists, experts, and doctors (Chinn and Hasell 2023; Eberl, Huber, and Greussing 2021; Merkley 2020; Mede and Schäfer 2020; Oliver and Rahn 2016). Though there are many reasons

why people may come to hold anti-establishment views, from personality traits, to economic resentment and rising political corruption (Piazza 2024; Uscinski et al. 2021). Researchers have begun to explore the relationship between anti-establishment views and social media use, including both political and apolitical social media (Chinn and Hasell 2023; Uscinski et al. 2021). This research suggests that individuals with anti-establishment views may be drawn to extremist social media content that promotes anti-establishment content, like conspiracy theories (Uscinski et al. 2021; Mitra, Counts, and Pennebaker 2021), but also that the incentive structures of social media platforms may cultivate and reward content that promotes anti-establishment views (Chinn, Hasell, and Shao 2026; Hasell and Chinn 2023; Starbird, DiResta, and DeButts 2023; Tripodi, Garcia, and Marwick 2024). There is some evidence of research conducted by platforms like Facebook and Twitter themselves finding that their platforms tend to reward and amplify political content that promotes anti-establishment views (Milmo 2021). However, there is little research that systematically examines how much content posted in social media may express anti-establishment views in seemingly apolitical contexts.

AES: We operationalize anti-establishment sentiment (AES) as expressions of distrust or criticism of public serving institutions for purposefully working against the interests of the American people (the People) (Droste 2021; Mudde 2004). We then focus on examining when, how, and how often AES occurs on social media. To be precise, we address the following research questions:

- What is the relationship between post topic and AES content?
  - How does the use of AES differ by post topic on TikTok?
  - How do users engage with content with AES and without it and how does that engagement vary across topics?
  - Which institutions are most likely to be the target of AES across topics?
  - Is the relationship between post topic and AES content similar across social media platforms?
- How prevalent is AES content on the For You Page?
- How do videos with AES content use linguistic style to create divisions between the People and elites?

Our first set of questions concerns how users interested in different topics may encounter AES in social media. AES research typically focuses on a given community, from healthcare (Chinn, Hasell, and Shao 2026), to conspiracy (Samory and Mitra 2018). We would like to systematically understand how the topics one is interested in influences their exposure to AES on TikTok, as well as how content creators in different genres are incentivized to create (or not) AES content, and the kinds of institutions they target. In order to address these questions we collect a core dataset of 26,783 videos using the TikTok Research API. We then develop a training program to annotate the data with the help of crowdsourced labelers. Finally, we train supervised learning models to annotate the remaining unlabeled data. Using the final human+machine labeled data, we address the questions above which depend on the relationship between post topic and AES content on TikTok.

Additionally, to understand whether the relationship between topic and AES prevalence generalizes to other platforms, we collect an additional dataset from YouTube. We sample and manually label a select number of posts from the three topics of conspiracy, finance, and wellness collected with the YouTube API. Finally, in order to understand how a user who is not systematically interested in a single topic may encounter AES content we collect data from the TikTok For You Page (FYP). To do so, we employ sock puppets to simulate interactions with the platform.

Finally, we are interested in how content creators on TikTok use linguistic style to create divisions between the People and elites. Such divisions are generally understood to be an important component of AES messaging (Droste 2021; Mudde 2004). Yet, it is not understood how this messaging might appear on TikTok, a platform which has been found to have a unique style (Herrman 2019).

Related Work

Despite increased interest in the role of anti-establishment views in democratic society, and the connection between those views and social media use, there has been little systematic analysis of the prevalence of AES in social media content, especially outside of political contexts or conspiracy theories. Research in the humanities and social sciences have shown that AES often appears in social media content. In some cases, this is because strategic, populist political actors are encouraging distrust and hostility towards established political actors and news media for their own purposes, and social media provides the affordances and capabilities for political influence outside of institutions (Tripodi 2022; Tripodi, Garcia, and Marwick 2024). In other cases, AES in social media content may be the result of everyday users navigating a crowded and cacophonous information environment at a time when most people are skeptical and distrustful of established, mainstream information sources (Chinn, Hasell, and Shao 2026; Cotter and Thorson 2022; The Knight Foundation 2023). As content creators compete for attention and influence, they are incentivized to establish themselves as credible sources of information, often positioning themselves as aligned with the people and in opposition to established institutions and industries (Chinn, Hasell, and Shao 2026; Hasell and Chinn 2023). However, the extent to which content may actually reflect AES remains unclear.

Within the computational social science community there has been work on detecting conspiracy theories (Diab, Nefriana, and Lin 2024), on understanding conspiracy talk as collective sensemaking (Kou et al. 2017), and on detecting, describing and understanding conspiratorial language (Steffen et al. 2023; Samory and Mitra 2018; Starbird 2017). This work tends to take conspiratorial discussions as a starting point, inspecting conspiracy theories in online spaces directly. There has also been work looking at populist rhetoric and political extremism and how the political establishment is portrayed negatively within such language (Grover and

Mark 2019; Jungherr, Posegga, and An 2022). However, anti-establishment views encapsulate far more than politics and government, as fears about hidden agendas of powerful actors pervade many industries, including medicine, banking, and agriculture, among others. For example, recent work has shown that there is a relationship between attitudes towards vaccinations and trust in institutions (Mitra, Counts, and Pennebaker 2021).

Our study examines how frequently such anti-establishment sentiments appear in social media content and whether such content may be encouraged by users via social media engagement, including likes, comments, and shares. We specifically examine two topic areas, finance and wellness, that are popular in social media generally, but are not necessarily associated with AES. We compare these to the area of conspiracy, which generally is considered to be a topic with high amounts of AES.

Finally, we analyze a sample of posts which were collected in Fall 2024 with the use of "sock puppets" (Bandy 2021). These posts from the FYP give us a glimpse of a random sample of Tik Tok content and provide an estimate of the AES content a typical user may encounter across topics.

A Computational Approach for Detecting AES

We contribute a novel study of AES on the social media platform TikTok, which is an increasingly popular platform in the U.S., especially among young people (McClain 2024; Faverio 2024). Though mostly used for entertainment, of those under 30, 50% now use TikTok as a source of news and information and 75% rely on TikTok for product reviews and recommendations by content creators (McClain 2024; Faverio 2024). This highlights that TikTok is increasingly seen as a useful source of information on a range of topics among young people.

To examine our research questions, we first describe our conceptualization of AES. Next, we translate this into a supervised learning task. We then describe our process for obtaining human annotations of this concept. Finally, we apply machine learning to annotate the remaining data and assess the performance of the machine learning method.

> **Definition 1** *Fundamentally, an anti-establishment view expresses that institutions are working against the interests of the American people. An anti-establishment view is any comment that expresses dislike, distrust, or criticism of a mainstream organization or industry in America related to a public serving institution (i.e., the government, NASA, the news media, universities, the EPA, etc.), health (doctors, scientists, the pharmaceutical industry, etc.), or finance (the stock market, banks, credit cards, financial industry, etc.) for working against the people.*

Conceptualization of AES We describe AES as shown in Definition 1. We presented this definition to human annotators when they are initially trained to perform the annotation. The definition we provide is more specific than typically academic definitions of anti-establishment views discussed above as we wanted to provide clear guidance to the annotators that reflect the types of messages that appear in social media.

Areas of focus

We focus on three topics: conspiracy, finance, and wellness. Our first topic *conspiracy*, is a natural topic to study for AES content because conspiracies often posit that small groups of powerful elites have organized events or suppressed knowledge in a way that disadvantages ordinary people (Douglas et al. 2019). We expect that this topic will serve as an upper bound of how much AES content we may see within a given topic. We note that not all conspiracy content is related to AES – often discussions of conspiracies are focused on their inconsistencies, absurdities, or debunking them. In contrast, we do not expect AES to be as common in finance and wellness. However, prior research suggests that these topics will express some amount of AES.

The next topic, *finance*, was selected because it is a popular topic on which people seek information and advice on social media, while related industries like the stock market, real estate, and banking are often criticized as being rigged against or unfair to ordinary people. We expect much of the finance content to be focused on utilitarian information, tips, and advice, as almost 80 percent of the Millennial and Gen Z demographic rely on social media for financial advice (Rose 2023). However, given rising income inequality and the sense that the rich have too much political power (Wike et al. 2025), it is also likely that some of this discussion contains AES. For example, crypto-currencies are often celebrated for operating outside of elite institutions and regulatory systems that "favor" the wealthy.

We examine *Wellness* content as it is a multi-trillion-dollar industry that has become highly popular online (Callaghan et al. 2021). Much of the wellness content in social media is devoted to selling products and services in efforts to achieve better physical and mental health outcomes, but it can also contain AES (Chinn, Hasell, and Hiaeshutter-Rice 2023). Wellness often criticizes western medicine and the pharmaceutical industry as wanting to keep people sick to profit off disease in ways that are often anti-establishment (Baker 2022). Content creators have an incentive to promote their own expertise over the advice of doctors and other experts to build their followings, and often position intuition and common sense over medical expertise (Carrion 2018).

We selected these topics as exemplars of the kinds of areas on social media where content creators build a brand around sharing advice based on their own experience, research, and expertise. Our work provides a reference point for future work, by establishing a baseline of AES content in online communities with incentives to express AES or not.

Problem Formulation

Our goal is to solve the task of detecting the presence of AES (labeled according to Definition 1). Let x be a social media post which can be described by a multimodal feature space. For example, x, may be described by textual or visual features. We would like to determine if x expresses AES.

Given a dataset of $\mathbb{D} = \{(x_1, y_1), \ldots, (x_N, y_N)\}$ where y_i can be described by a binary label, and x_i belongs to a

multimodal feature space $\mathcal{X}$ which depends on the available data, our goal is to learn a function $f : \mathcal{X} \rightarrow \{0, 1\}$.

Concept elicitation with human annotation

To detect the presence of AES, we employ human annotators. We utilized Prolific (Prolific 2025), a research-focused online platform, to recruit participants for the annotation task. These annotators must first go through a training program before they are asked to perform data annotation. We conducted both the training program and the main annotation through Qualtrics (Qualtrics 2025), an online survey tool.

Our survey was approved by an internal IRB. All annotators gave informed consent to both the training and the annotation task. The training and annotation contained warnings that annotators may encounter explicit content. Annotators were given the option to not review any content that they preferred not to review.

Training: The training program includes: (1) an initial description of the task, (2) examples of posts which do and do not include the target concept, (3) test questions with feedback where annotators can practice their understanding of the concepts, and (4) a final assessment. In the initial description of the task we explicitly define the concept of AES and provide examples of common institutions.

We then ask them to determine if a video mentions an institution or not and provide feedback about their responses. In particular we ask if any of the following institutions are mentioned in the post:

- The federal government or government agencies (e.g. NASA)
- the news media (e.g. Washington Post or Fox News)
- the banking industry (e.g. Wall Street)
- the pharmaceutical industry or medical research (e.g. Big Pharma)
- scientific research or organizations (e.g. the EPA or an Academic University)
- other institution such as agriculture, the oil industry, or politicians

Next, we show example posts and ask participants to determine whether the post expresses AES. We then provide feedback on the selection. For example, consider a video with the following phrase "If you wear a NASA t-shirt and it's not a parody. Man they got you. You're in a trance dog.". We say that this post does contain AES. If the participant was correct we reiterate why it expresses AES (negative sentiment towards the institution NASA), if they are incorrect we point out what they may have missed (NASA is a federally funded entity and we would consider it an institution. The poster is generally negative about NASA). We also include examples of posts which do not contain AES, and again provide feedback depending on the participant's selection.

After participants have interacted with examples around whether a post mentions an institution and whether a post expresses AES, we turn to examples around comments. Here, we are interested in (1) whether a comment appears to agree with the post, (2) whether a comment expresses AES.

We describe agreement as: the comment praises the video, expresses gratitude towards the content creator, or criticizes those who disagree with the video. In contrast, we describe disagreement as: the comment criticizes the content creator's viewpoint, makes fun of the video, questions the credibility of the video, or dismisses the content creator's values. We follow the same format as above, except that here participants are shown both a post and a comment made in response to the post. They are then asked to determine agreement with the post and the AES content of the comment. They are given feedback about each example they complete.

The final step in the training is the assessment. The assessment contains 16 questions. In order to be eligible for the annotation task, annotators must achieve a score of 75% on the assessment.

Annotation: We structure the task around pairs of videos and comments. Before performing the annotation, an annotator must pass a brief assessment of 4 questions, to demonstrate that they have retained the information from the training. That is, an annotator is shown a video and then shown a comment posted in response to the video. They are then asked the questions shown in Table 1.

Machine annotation

After collecting labels from human annotators and aggregating the labels to arrive at a single label per example, we next employ machine learning to label the remaining instances in the dataset.

Utilizing categorical information We also experiment with an additional feature representation for this task. The videos are collected to be representative of three distinct categories of content: conspiracy theories, finance, and wellness. As each category may have distinct linguistic signals which suggest AES we may benefit from representing categorical information.

In both the baseline and deep learning settings we incorporate categorical information directly. That is, in the baseline setting we concatenate a one-hot feature vector which encodes a video's category with the text-based feature vector. In the deep learning setting, we train an embedding layer which also takes a category vector as input.

Empirical Evaluation

We collect posts from the social media platforms TikTok and YouTube. First, we describe our data collection process. Next, we evaluate both the human and machine annotations.

Dataset

We collect $26,783$ video posts with $206,350$ comments using the TikTok Research API. Our goal was to collect posts oriented around the three themes of finance, wellness, and conspiracy theories in recent years. Thus, we set the following parameters in the API requests:

- **Key words**: the post could contain any of the keywords from the sets of FINANCE, WELLNESS, CONSPIRACY.

Question	Scale
Video Annotation	
Do you think the person in the video is expressing anti-establishment views?	4-point (from *Yes, they are **definitely** expressing anti-establishment views* to *No, they are **definitely not** expressing anti-establishment views*)
Comment Annotation	
Do you think the comment agrees with the information in the video?	5-point (from *Yes, I think this comment **definitely agrees** with the information in the video* to *No, I think this comment **definitely does not agree** with the information in the video*)

Table 1: Schematic of annotation task.

	CONSPIRACY	FINANCE	WELLNESS	Total
Number of videos	8,439	8,956	9,388	26,783
Number of comments	96,217	62,154	47,978	206,349

Table 2: Distribution of videos and comments across three categories in the datasets: CONSPIRACY, FINANCE, and WELLNESS. The primary TikTok dataset comprises a total of 26,783 videos and 206,349 comments.

- **Timeline:** the post was made between January 1, 2022, and December 31, 2023 for FINANCE, WELLNESS, CONSPIRACY.
- **Location:** the post was made in the United States.

Where CONSPIRACY contained the phrases: *"conspiracy", "flatearth", "propaganda", "illuminati"*, FINANCE contained the phrases: *"finance","stocks", "crypto", "realestate"*, and WELLNESS contained the phrases: *"wellness","health", "selfcare", "fitness"*. To derive the phrases in each set we began with three umbrella terms: *conspiracy*, *finance*, and *wellness*. Then we created a seed set of a few thousand posts which contained each umbrella term. We then chose common words within the seed set to add to one of the three sets of FINANCE, WELLNESS, or CONSPIRACY. When there was a word which appeared to return posts unrelated to the target concept, we did not add it to the set. Overall this process produced the dataset shown in Table 2.

We describe the dataset at different stages of the analysis in Table 3. In the first phase (**Data Collection and Transcription**), we download videos from TikTok in the process described above. This dataset consists of only those videos which we were able to transcribe. Next, in the **Data Filtering** phase, we retain only those videos such that length of the transcription and the video description was at least 40 tokens, and the language of the text was detected as English[1]. In the **Data Annotation** phase, we select a sample of video-comment pairs, where each comment to be annotated contains at least 10 tokens. Finally, we clean this dataset to only contain high quality annotations (**Annotated Data Filtering**). To do so, we only considered annotations with at least three labels from the best annotators (see Human Annotation - Empirical Details). The **Gold Set** dataset is a subset of videos which the domain expert on the team annotated in terms of AES.

[1]To detect the language of the text we used the spaCy package (Honnibal et al. 2020)

Stage	Videos	Comments
Data Collection and Transcription	26,783	206,350
Data Filtering	14,261	129,996
Data Annotation	816	890
Annotated Data Filtering	616	N/A
Gold Set	103	N/A

Table 3: For each state of data processing we display the number of videos and comments available. We did not consider the annotation of comments when filtering the annotated data, nor did we annotate comments in the Gold Set creation.

Human Annotation - Empirical Details

In the annotation task, annotators were asked to label 10 video-comment pairs. Recall that, before starting the task, they were required to pass an assessment consisting of 4 questions, and they needed to achieve a perfect score to qualify for the task. Moreover, to maintain high-quality annotations, we included two attention checks within the task. Annotators were required to pass both attention checks to successfully complete and submit their annotations. Ultimately, 39 annotators provided valid annotations, meeting all the requirements. These annotators were reimbursed at an average rate of $21.90 per hour. Each video-comment pair was independently annotated by three different annotators.

To validate the human annotation we annotate a small gold set of 103 videos. We then inspect the precision, recall, F1 Score, and accuracy of the annotations under different aggregation schemes. In Table 4 we see that the best results are obtained with the Dawid Skene aggregation method. Additionally, we see that we obtain reasonable performance with a precision of 0.556. Thus, for the remainder of the paper we use the aggregate labels as determined by Dawid Skene.

Machine Annotation - Implementation Details

We treat our dataset as a text dataset. That is, for each video we produce a transcript using WhisperX (Bain et al. 2023). We then concatenate the video description to the video transcript to create a single text document for each video. Given

Aggregation Method	Dawid Skene (Dawid and Skene 1979)	MACE (Hovy et al. 2013)	Majority Vote
Precision	0.556	0.528	**0.562**
Recall	**0.909**	0.864	0.818
F1 Score (binary)	**0.690**	0.665	0.667
F1 Score (macro)	**0.784**	0.760	0.774

Table 4: Performance comparison of different aggregation methods, evaluated using a gold standard set of size 103. Metrics include Precision, Recall, Binary F1 Score, and Macro F1 Score. In general, Dawid Skene does the best among these three methods and hence we used this aggregation method for the remainder of the paper.

this text dataset, we employ language models to annotate the remainder of the dataset.

We would like to exploit the general knowledge stored in language models which have been trained on large quantities of text. Thus, we use BERT (Devlin et al. 2019), DistilBERT (Sanh et al. 2019), RoBERTa (Liu et al. 2019) and DistilRoBERTa (Sanh et al. 2019) as state-of-the-art pretrained language models (PLMs). Additionally, we compare these approaches to two baselines: SentenceBERT (Reimers and Gurevych 2019) plus SVM and SentenceBERT (Reimers and Gurevych 2019) plus LightGBM (Ke et al. 2017).

As the structure of AES content will likely vary with the category, we experiment with the inclusion of category specific information. The way we incorporate such information depends on the model. For the SVM and LightGBM models, we one-hot encoded the category information and directly included it as a feature for training the models.

For the PLMs, we pass texts to a PLM which outputs textual representations. To incorporate categorical information, we generate categorical embeddings. That is, we train an Embedding layer which takes a category vector as input and outputs a dense embedding. These category embeddings are concatenated with the PLM-derived representations. The combined feature vector is then passed through a fully connected layer that maps the concatenated representations to a binary label space. The entire model, including the PLM, category embeddings, and the classification layer, is trained jointly in an end-to-end fashion to optimize performance on the binary classification task. This approach enables the model to effectively integrate semantic and categorical information for improved prediction accuracy. The combined textual and categorical deep learning models are implemented in PyTorch (Paszke et al. 2019).

For each case, we select hyperparameters by performing a stratified 5-fold cross-validation grid search on the training set, consisting of a total of 500 videos. Each grid search is conducted using three distinct random seeds, and the hyperparameters achieving the best performance across these seeds are selected. The results on the held-out test set of 116 videos, averaged across the three random seeds, are reported in Table 5.

The inclusion of category specific information was helpful for this task. We inspect the binary F1-score, as an informative measure in the setting with class imbalance. The best F1-score of 0.728 is achieved by RoBERTa with the inclusion of category-specific information.

Training Details: For the RoBERTa model in the Text + Category setting, the final hyperparameters are as follows: PLM: roberta-large, learning rate: $1 \cdot 10^{-5}$, epochs: 40, size of the categorical embeddings: 32, and class weights for positive and negative labels: 0.35 and 0.65, respectively. All models were trained on an AMD EPYC 7763 64-Core CPU and three NVIDIA RTX A5000 GPUs.

Research Questions and Findings

Next, we return to our original research questions. First, we annotate the entire dataset using the best machine learning model found using the validation set and trained on the human annotations: RoBERTa with categorical information. Then, we use this entire dataset of both human and machine annotations to answer the research questions.

How does the use of AES differ by post topic on TikTok? We see in Figure 1 that the prevalence of AES sentiment varies widely across the categories. Results show the greatest proportion of AES is found in the CONSPIRACY category, and we see little AES content in the categories of FINANCE (less than 5%) and WELLNESS (∼1%). The high proportion of AES content in CONSPIRACY serves to validate our approach.

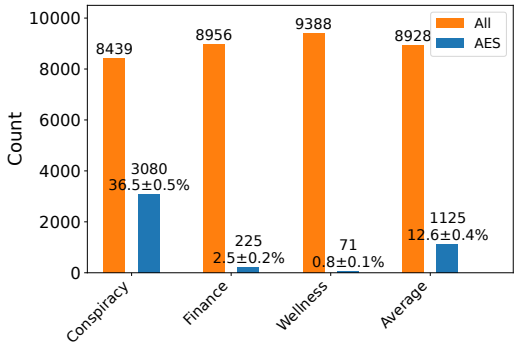

Figure 1: Total video counts are shown above the bars. CONSPIRACY has the highest AES proportion (45.1%), followed by FINANCE (4.3%) and WELLNESS (1.3%), with an overall average of 16.1%. Confidence intervals across 3 seeds.

How do users engage with content with AES and without it and how does this engagement vary by topic? In Figure 2, we see differences in some forms of engagement depending on whether a post contains AES or not. For example, in the area of FINANCE AES content receives slightly more comments and shares, than non-AES content. In the

Pure Text						
Metrics	**SentenceBERT + SVM**	**LightGBM**	**BERT**	**DistilBERT**	**RoBERTa**	**DistilRoBERTa**
Precision	$0.619_{0.000}$	$\mathbf{0.833_{0.000}}$	$0.476_{0.006}$	$0.501_{0.011}$	$0.537_{0.051}$	$0.520_{0.010}$
Recall	$\mathbf{0.765_{0.000}}$	$0.441_{0.000}$	$0.580_{0.026}$	$0.570_{0.009}$	$0.755_{0.067}$	$0.614_{0.023}$
F1 Binary	$\mathbf{0.684_{0.000}}$	$0.577_{0.000}$	$0.521_{0.007}$	$0.533_{0.006}$	$0.617_{0.019}$	$0.562_{0.006}$
Accuracy	$0.793_{0.000}$	$\mathbf{0.810}_{0.000}$	$0.731_{0.004}$	$0.747_{0.007}$	$0.721_{0.046}$	$0.758_{0.006}$
Text + Category						
Metrics	**SentenceBERT + SVM**	**LightGBM**	**BERT**	**DistilBERT**	**RoBERTa**	**DistilRoBERTa**
Precision	$0.553_{0.000}$	$0.600_{0.000}$	$0.585_{0.011}$	$0.562_{0.032}$	$\mathbf{0.659_{0.009}}$	$0.555_{0.042}$
Recall	$0.765_{0.000}$	$0.265_{0.000}$	$0.676_{0.017}$	$0.657_{0.020}$	$\mathbf{0.814_{0.020}}$	$0.725_{0.043}$
F1 Binary	$0.642_{0.000}$	$0.367_{0.000}$	$0.627_{0.013}$	$0.604_{0.022}$	$\mathbf{0.728_{0.013}}$	$0.624_{0.012}$
Accuracy	$0.750_{0.000}$	$0.733_{0.000}$	$0.764_{0.008}$	$0.747_{0.020}$	$\mathbf{0.822_{0.008}}$	$0.741_{0.028}$

Table 5: Comparison of model performance across two experimental setups: one using pure text features and the other incorporating text with category embeddings. Metrics evaluated include average Precision, Recall, F1 Binary, and Accuracy, along with their respective standard errors, derived from experiments conducted with three distinct random seeds. In the pure text setup, SentenceBERT + SVM achieves the highest F1 Binary (0.684) and Recall (0.765), while LightGBM excels in Precision (0.883) and Accuracy (0.810). With category embeddings, RoBERTa demonstrates the strongest performance across all metrics.

area of WELLNESS it receives more comments per post, but far fewer shares per post. In all three categories it receives fewer likes per post.

In the annotation task, annotators are asked to determine if a comment made in response to a post agrees with the content of the post, disagrees with the content of the post, or seems to be irrelevant. We now use those labels to understand if viewers of posts with AES tend to agree or disagree with the content in the posts. Here, we take the majority vote for each comment label, and if there is a three-way tie, or if the majority vote assigns the irrelevant label, we mark the agreement of the comment as unclear.

In Figure 3, we analyze a subset of comments made in response to TikTok posts in the CONSPIRACY category (the only category with sufficient comment annotations to warrant analysis). Here, we see that in aggregate commenters are agreeing with AES content as much, if not more so, than non-AES content.

Which institutions are most likely to be the target of AES? A particularly important question towards understanding anti-establishment sentiment is which institutions are the target of such sentiment. To address this question we hand coded a sample of 50 posts which were predicted to contain AES.

In particular we coded each post as to whether it contained the following institutions:

- **US Government:** A mention that the US government is not to be trusted.
- **Politicians:** A mention that politicians are not to be trusted/are corrupt.
- **US Healthcare:** A mention that there is something fundamentally wrong/untrustworthy about the US Healthcare system.
- **Big Pharma:** A mention that there is something fundamentally wrong/untrustworthy about pharmaceutical companies.
- **Big Banks:** A mention that there is something fundamentally wrong/untrustworthy about financial institutions.
- **NASA:** A mention that NASA is corrupt and/or lies to the American public.

In Figure 4, we offer a cursory glimpse into which institutions are discussed in these posts across the different categories. We manually code 50 posts, thus we don't present this analysis as representative, but as a way of understanding how the categories likely differ. For example, US Healthcare is mentioned only in Wellness posts, but posts referring to the US Government appear both in conspiracy and finance posts. When a particular US Government body was mentioned (like NASA) we noted this, and results suggest that references to the government in the abstract are more common than accusations targeted at specific bodies.

Is the relationship between post topic and AES content similar across social media platforms? To address this question we use the same keywords for each of the three topics and collect $\sim$80 posts using the YouTube API for each set of keywords for a total of 1013 posts. We set the region code to the United States, the language to English, and the timeline to be between January 1, 2023 and December 31, 2024. This sample allows us to see how AES content differs from TikTok to YouTube for the same set of topics.

We find that the ranking of the three topics in terms of AES prevalence is the same. That is, 20% of the conspiracy posts, 4% of the finance posts, and less than 1% (.06%) of the wellness posts contain AES. This shows that users interested exclusively in wellness would be least likely to encounter AES on either platform. Moreover, like on TikTok, users primarily interested in finance would be slightly more likely to encounter AES than users interested in wellness. Our sample also suggests that users who exclusively use YouTube would be less likely to encounter AES in conspiratorial content than those who exclusively use TikTok. This may be a result of YouTube's policies to suppress problematic content, (Buntain et al. 2021).

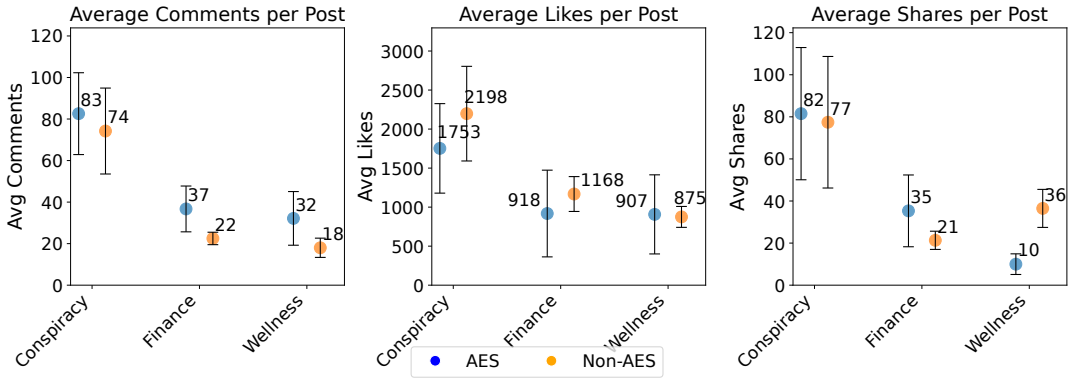

Figure 2: Engagement metrics (Comments, Likes, Shares) per post, categorized by AES (Anti-Establishment Sentiment) and Non-AES content across CONSPIRACY, FINANCE, and WELLNESS topics. AES posts show lower engagement in all metrics for CONSPIRACY; AES posts show higher engagement in Comments and Shares but lower engagement in Likes for FINANCE; and AES posts show higher engagement in Comments but lower engagement in Likes and Shares for WELLNESS.

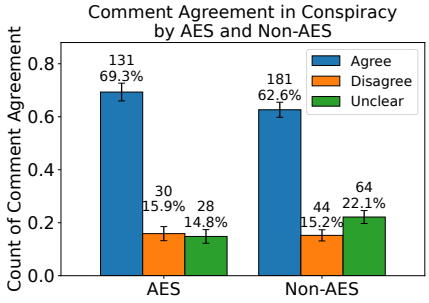

Figure 3: Distribution of comment agreement in the CONSPIRACY category for AES (anti-establishment) and Non-AES labels. AES comments show a higher agreement rate (69.3%) compared to Non-AES comments (62.6%). This suggests that engagement with conspiracy content is not-driven by users refuting conspiracy theories.

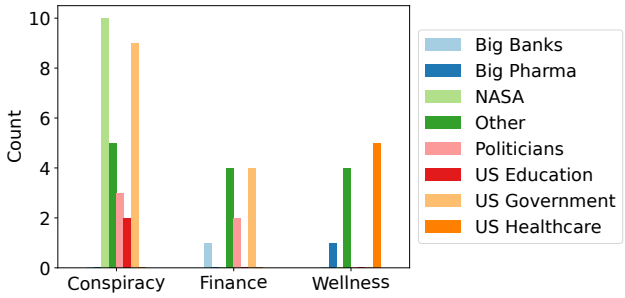

Figure 4: We see that the institutions which are the targets of AES vary by category. The most common target across categories is the US Government, while the Conspiracy videos with many Flat Earth videos target NASA, the Finance videos have diffuse targets, and the Wellness videos target the US Healthcare system.

For You Page Data We additionally collect data by mimicking human users who would log into the FYP on TikTok. To do so we create 48 sock puppet accounts. These accounts were created in the months of September and October 2024 and browsed content from October 2024 to December 2024 (35 days on average). To mimic a user who may be interested in a few topics, and the news, when each account is created it is randomly assigned 6 lifestyle accounts and 6 news channel accounts to follow (shown in the Appendix in Table 8 and Table 9). Each day each account logged on to their FYP at a randomly assigned time between 7:00 am and 9:00 am EST. They then logged on again exactly twelve hours later. In each session an account paused scrolling to watch each video with a 50% probability.

Unlike the other datasets this dataset is not filtered to particular keywords. It thus offers a more generalized view into how often AES content is shown to a user who only visits the FYP, rather than seeking topic-driven content.

All together, this procedure produced 37,012 videos. The research team annotated 349 of these videos for AES. We then trained a RoBERTa model as that performed the best in the other experiments. With RoBERTa, we achieved a binary F1 of 0.822, a precision of 0.889, a recall of 0.833, and a accuracy of 0.990. We thus proceeded to infer the labels of the remaining videos utilizing this trained model.

We found that less than 1% (0.48%) of all posts expressed AES. This again shows that the extent to which a user will be exposed to AES on TikTok depends on their interests. While a user who periodically browses the FYP may see very little AES content, a user interested in conspiracy theories will see much more.

How do videos with AES content use linguistic style to create divisions between the People and elites? Returning to our definition of AES as an expression of a perceived power struggle between the People and elites (or the establishment) (Uscinski et al. 2021), we would like to see how this tension surfaces through the linguistic style of content creators on social media. We propose that such divisions will occur through three primary themes:

- Authenticity: Creators of AES content may portray the people as having real authenticity and elites as lacking connections to the lives and circumstances of ordinary individuals' lived experiences.
- Authority: Creators of AES content may use a style which vests themselves with authority and the elites as lacking the real authority of the people.
- Morality: Creators of AES content may use a moralistic style, speaking to the morality of the people and the amorality of elites. Religious themes will be evoked with the people being on the side of good faith and the elites being portrayed as lacking faith.

To explore these three themes we employ the most recent version of the Linguistic Inquiry and Word Count (LIWC2) software (Boyd et al. 2022). LIWC analyzes text by calculating the proportion of words that fall into predefined linguistic categories. For each of the three themes we inspect specific cues which signal the use of the theme. Table 6 shows these themes and the factors and variables assigned to them.

Authenticity reflects the extent to which language is personal and self-revealing, as opposed to detached and guarded (Boyd et al. 2022). Language with high authenticity scores is characterized by a greater use of I-words, present-tense verbs, and relativity words (e.g., *old*, *far*, *here*), and a reduced use of third-person pronouns (e.g., *she*, *he*) and discrepancy words.

Theme	Factors	Variables	Labels	Mean	SE
Authenticity	Persuasive Tone	Authenticity	Non-AES	48.7	0.25
			AES	43.8	0.47
	In-out-group	First-person singular	Non-AES	3.91	0.03
			AES	2.34	0.05
		First-person plural	Non-AES	0.81	0.01
			AES	1.39	0.03
		Third-person plural	Non-AES	0.49	0.01
			AES	1.39	0.03
Authority	Relevance	Clout	Non-AES	56.3	0.26
			AES	62.1	0.45
		Power-related	Non-AES	0.81	0.01
			AES	1.98	0.04
	Gender	Male reference	Non-AES	0.62	0.01
			AES	0.82	0.02
		Female reference	Non-AES	0.53	0.01
			AES	0.29	0.02
Morality	Relevance	Religion-related	Non-AES	0.31	0.01
			AES	0.79	0.04
		Death-related	Non-AES	0.09	4e-3
			AES	0.18	9e-3

Table 6: Summary of linguistic cue measurements across different variables and labels. The table categorizes linguistic features into four main factors: Persuasive Tone, Relevance, In-Group vs. Out-Group, and Gender. For each variable within these factors, the mean and standard error (SE) values are reported separately for Non-AES and AES labels.

Authority reflects the extent to which language asserts epistemic knowledge through certainty and dominance. We use the LIWC variable Clout to measure authority in the language of a post. A higher Clout score indicates a more powerful and confident language style, which is characterized by a greater use of we words and social words, alongside a reduced use of I words, negations (e.g., *no*, *not*), and swear words. Power-related words, such as *own*, *order*, *allow*, and *power*, reflect themes of dominance, control, and influence.

Morality reflects the extent to which language is used to position events or ideas as in the interest of society or the good of the people, as opposed to harm or evil. We use the LIWC variable Religion-related words, such as *god*, *hell*, and *church*, indicate discussions of spirituality, beliefs, or religious practices, often reflecting cultural or moral perspectives. Death-related words, including *die*, *kill*, and *coffin*, highlight themes of mortality, loss, and existential concerns, which can signal emotional intensity or warnings.

The linguistic analysis reveals distinct patterns between AES and non-AES content across key dimensions. When it comes to Authenticity, we see that non-AES content demonstrates higher Authenticity (48.7 vs. 43.8), suggesting greater personalization through first-person narratives and present-tense language, a strategy linked to trust-building in lifestyle content (Hasell and Chinn 2023). In-out-group dynamics further highlight differences: AES posts favor collective identity (first-person plural: 1.39 vs. 0.81; third-person plural: 1.39 vs. 0.49) over individual perspective (first-person singular: 2.34 vs. 3.91), reinforcing the "people vs. elites" dichotomy central to anti-establishment rhetoric (Mudde 2004). This aligns with Klein's findings on social media posts promoting conspiracy theories, which similarly employ in-group and out-group language to create division, leverage emotional appeals to provoke and engage audiences, and reference religious doctrines as a means of validation (Klein 2023). By strategically using these rhetorical devices, AES posts not only reinforce the "people vs. elites" dichotomy but also position themselves as credible and authoritative voices, effectively garnering trust and influence among their audiences.

When it comes to Authority, the Persuasive Tone metrics show AES posts exhibit significantly higher Clout scores (AES: 62.1 vs. Non-AES: 56.3), indicating a more authoritative and collective language style. This aligns with findings that anti-establishment creators position themselves as alternative authorities by adopting confident rhetoric to challenge institutional expertise (Chinn, Hasell, and Shao 2026).

When it comes to Morality, we see that religious-related terms are more prevalent in AES content (religion: 0.79 vs. 0.31), reflecting that some people tend to find reference from religious documents. (e.g., "The earth is flat. You can find it in Bible."). Death-related terms are also more prevalent in AES content (death: 0.18 vs. 0.09), echoing that AES spreads fear and anger by association with death. (e.g., "Take a vaccine that could maybe make me die").

Gender-related language shows minor disparities in ways which may relate to authority and morality: AES posts use more male references (0.82 vs. 0.62) and fewer female references (0.29 vs. 0.53), potentially reflecting gendered stereotypes in conspiratorial narratives (e.g., framing male figures as dominant antagonists). Research indicates that conspiracy theories often employ gendered language, with a tendency to reference male figures more frequently than female ones. Work analyzing gender representations in conspiracy discourse found that such narratives often reinforce connections between religiosity and masculinity, while relying on biological gender essentialism to define femininity (Fleckenstein 2025). The results from our study suggest that male

figures are more prominently featured, potentially as dominant antagonists in AES posts.

In summary, we see that AES content creators use linguistic style to build their own authority and can draw on themes related to morality. AES content creators score lower on the authenticity variable than non-AES content creators. While we hypothesize that these content creators communicate authenticity in different ways, this finding also creates opportunities for future study.

Discussion

This study systematically examines the prevalence of anti-establishment sentiment on TikTok across a range of topics that are not explicitly political. Prior research has shown that social media platforms tend to reward and amplify political content that promotes anti-establishment sentiment (Milmo 2021; Starbird, DiResta, and DeButts 2023; Tripodi, Garcia, and Marwick 2024; Zadrozny 2021), and other work has suggested that even apolitical content on social media may facilitate and encourage anti-establishment views (Chinn, Hasell, and Shao 2026; Hasell and Chinn 2023) as influencers and content creators position themselves as voices of expertise and authority in social media (Wellman 2024). Our results suggest that AES in posts about finance and wellness is relatively uncommon. We found AES in an estimated 4% of all posts we collected in the domain of FINANCE and 1% of all posts we collected in WELLNESS. In contrast, we find that 45% of all CONSPIRACY posts contained AES. Overall, this suggests that anti-establishment sentiment is not widespread in apolitical content on TikTok.

However, such posts may generate more comments and shares than other types of content. For example, in FINANCE and WELLNESS, our results show that AES content is more likely to be commented on, and in FINANCE such content is more likely to be shared. Though we also note that across all three categories, including CONSPIRACY, AES content is liked less often. In regard to the comments, in Figure 3, we find generally that comments agree with the posted content. Upon manual inspection we do see occasional posts which will ridicule for example, flat-earth theories, but in aggregate, in our sample of labeled data we see that comments tend to agree with the original post.

Future work should consider both how anti-establishment sentiment might appear in a wider range of topics and the potential long-term effects of encountering such content. Though only a small portion of posts contained AES, there remain questions as to how influential AES content might be when users are repeatedly exposed over time.

We propose including a categorical representation with the textual representation of each post and find this approach to be useful. When such information was included, we see that RoBERTa achieves the best performance overall. This is unsurprising, as this is the one of the larger models we experiment with. Future work could focus on a single category, e.g. wellness, and train models specifically for that domain.

Finally, we would like to note that the concept of AES is subtle. Annotating the Gold Set required much deliberation across the research team to ensure that our annotations were consistent and in line with what crowd workers could do. We see a line of future work seeking to understand how this term can best be measured as it evolves.

Limitations

One significant limitation of this work is that we did not investigate the number of views for each post. However, this was not an analytical oversight, but an empirical decision. We found that the view data returned by the TikTok API was unreliable. That is, there were posts with 0 views but more than 0 likes. We researched this discrepancy and found a range of non-conclusive hypotheses. For example, it is possible that TikTok fails to reliably log these characteristics. It is also possible that TikTok manually sets the views of certain types of content or accounts to 0. However, we found no difference in the prevalence of this discrepancy across the three categories.

Ethical Implications There may be concerns that collecting posts made by individuals is a violation of privacy. However, the posts we accessed were all publicly available and we do not share any individual's data or content. For example, we do not publish data related to whether any individual post or content creator was labeled as anti-establishment. Additionally, all annotators were provided with informed consent and an institutional IRB reviewed the study and ruled it to be exempt.

Conclusion

We find that AES content does occur across three types of content on TikTok, conspiracy theories, finance, and wellness, to varying degrees. Such content is common in conspiracy content, but rare in the domains of finance and wellness. This result generalized to a sample of YouTube content as well. That is, we found that the relative prevalence of AES in these topics was the same on YouTube, while YouTube conspiracy content expressed less AES than conspiracy content on TikTok. We also considered a collection of posts from TikTok's FYP. Here, we found that AES is expressed in less than 1% of all content. Thus, users will encounter different levels of exposure to AES, depending on their interests and behaviors. A user who browses the FYP with limited interaction will likely encounter very little AES content, and a user who seeks conspiratorial content will likely encounter it regularly. However, better understanding the implications of such content requires further testing of the effects of exposure to AES in social media and examination over longer periods of time. With anti-establishment views increasing, and the increased prominence of political figures who hold anti-expert and anti-establishment views (Zadrozny 2025), it may be that such content will become increasingly prevalent.

References

Ad Fontes Media. 2025. Media Bias Chart. Accessed: 2025-04-23.

Armaly, M. T.; and Enders, A. M. 2024. Who supports political violence? *Perspectives on Politics*, 22: 427–444.

Bain, M.; Huh, J.; Han, T.; and Zisserman, A. 2023. WhisperX: Time-Accurate Speech Transcription of Long-Form Audio. *INTERSPEECH 2023*.

Baker, S. A. 2022. Alt. Health Influencers: how wellness culture and web culture have been weaponised to promote conspiracy theories and far-right extremism during the COVID-19 pandemic. *European Journal of Cultural Studies (EJCS)*, 25: 3–24.

Bandy, J. 2021. Problematic machine behavior: A systematic literature review of algorithm audits. In *Proceedings of the ACM Conference on Human-Computer Interaction (CSCW)*.

Barr, R. R. 2009. Populists, outsiders and anti-establishment politics. *Party politics*, 15: 29–48.

Boyd, R. L.; Ashokkumar, A.; Seraj, S.; and Pennebaker, J. W. 2022. The development and psychometric properties of LIWC-22. *Austin, TX: University of Texas at Austin*, 10(1-47): 6.

Buntain, C.; Bonneau, R.; Nagler, J.; and Tucker, J. A. 2021. YouTube recommendations and effects on sharing across online social platforms. *Proceedings of the ACM Conference on Human-Computer Interaction (CSCW)*.

Callaghan, S.; Lösch, M.; Pione, A.; and Teichner, W. 2021. Feeling good: The future of the $1.5 trillion wellness market. Accessed 2025-04-03.

Carrion, M. L. 2018. "You need to do your research": Vaccines, contestable science, and maternal epistemology. *Public Understanding of Science (PUS)*, 27: 310–324.

Chinn, S.; and Hasell, A. 2023. Support for "doing your own research" is associated with COVID-19 misperceptions and scientific mistrust. *Harvard Kennedy School Misinformation Review*.

Chinn, S.; Hasell, A.; and Hiaeshutter-Rice, D. 2023. Mapping digital wellness content: implications for health, science, and political communication research. *Journal of Quantitative Description: Digital Media*, 3: 1–56.

Chinn, S.; Hasell, A.; and Shao, A. 2026. What does it mean to "do your own research?" A comparative content analysis of DYOR messages in Instagram and Facebook posts about reproductive health, food, and vaccines. *New Media & Society*, 28: 567–588.

Cotter, K.; and Thorson, K. 2022. Judging value in a time of information cacophony: Young adults, social media, and the messiness of do-it-yourself expertise. *The International Journal of Press/Politics*, 27(3): 629–647.

Dawid, A. P.; and Skene, A. M. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28: 20–28.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the conference of the North American chapter of the association for computational linguistics (NAACL)*.

Diab, A.; Nefriana, R.; and Lin, Y.-R. 2024. Classifying Conspiratorial Narratives at Scale: False Alarms and Erroneous Connections. In *Proceedings of the AAAI Conference on Web and Social Media (ICWSM)*.

Douglas, K. M.; Uscinski, J. E.; Sutton, R. M.; Cichocka, A.; Nefes, T.; Ang, C. S.; and Deravi, F. 2019. Understanding conspiracy theories. *Political psychology*, 40: 3–35.

Droste, L. 2021. Feeling left behind by political decisionmakers: Anti-establishment sentiment in contemporary democracies. *Politics and Governance*, 9: 288–300.

Eberl, J.-M.; Huber, R. A.; and Greussing, E. 2021. From populism to the "plandemic": Why populists believe in COVID-19 conspiracies. *Journal of Elections, Public Opinion and Parties*, 31: 272–284.

Enders, A. M.; Diekman, A.; Klofstad, C.; Murthi, M.; Verdear, D.; Wuchty, S.; and Uscinski, J. 2023. On modeling the correlates of conspiracy thinking. *Scientific Reports*, 13: 8325.

Enders, A. M.; and Uscinski, J. E. 2021. Are misinformation, antiscientific claims, and conspiracy theories for political extremists? *Group Processes & Intergroup Relations (GPIR)*, 24(4): 583–605.

Faverio, M. 2024. A majority of U.S. TikTok users are there for product reviews and recommendations. Accessed: 2024-12-14.

Fleckenstein, K. 2025. Representations of gender in conspiracy theories: a corpus-assisted critical discourse analysis. *Critical Discourse Studies*, 22: 357–373.

Grover, T.; and Mark, G. 2019. Detecting potential warning behaviors of ideological radicalization in an alt-right subreddit. In *Proceedings of the AAAI Conference on Web and Social Media (ICWSM)*.

Hasell, A.; and Chinn, S. 2023. The political influence of lifestyle influencers? Examining the relationship between aspirational social media use and anti-expert attitudes and beliefs. *Social Media+ Society*, 9: 20563051231211945.

Herrman, J. 2019. How TikTok is rewriting the world. *The New York Times*. Accessed 2025-04-23.

Honnibal, M.; Montani, I.; Van Landeghem, S.; Boyd, A.; et al. 2020. spaCy: Industrial-strength natural language processing in Python.

Hovy, D.; Berg-Kirkpatrick, T.; Vaswani, A.; and Hovy, E. 2013. Learning whom to trust with MACE. In *Proceedings of the conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Jungherr, A.; Posegga, O.; and An, J. 2022. Populist supporters on Reddit: A comparison of content and behavioral patterns within publics of supporters of Donald Trump and Hillary Clinton. *Social Science Computer Review (SSCR)*, 40(3): 809–830.

Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; and Liu, T.-Y. 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems (NeurIPS)*.

Klein, E. 2023. *Loaded language and conspiracy theorizing*. Ph.D. thesis, Rensselaer Polytechnic Institute.

Kou, Y.; Gui, X.; Chen, Y.; and Pine, K. 2017. Conspiracy talk on social media: collective sensemaking during a public health crisis. *Proceedings of the ACM Conference on Human-Computer Interaction (CSCW)*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lyons, B. A. 2023. How orientations to expertise condition the acceptance of (mis) information. *Current Opinion in Psychology*, 54: 101714.

McClain, C. 2024. About half of TikTok users under 30 say they use it to keep up with politics, news. Accessed: 2024-12-14.

Mede, N. G.; and Schäfer, M. S. 2020. Science-related populism: Conceptualizing populist demands toward science. *Public Understanding of science*, 29: 473–491.

Merkley, E. 2020. Anti-intellectualism, populism, and motivated resistance to expert consensus. *Public Opinion Quarterly*, 84: 24–48.

Milmo, D. 2021. Twitter admits bias in algorithm for rightwing politicians and news outlets. *The Guardian*.

Mitra, T.; Counts, S.; and Pennebaker, J. 2021. Understanding anti-vaccination attitudes in social media. In *Proceedings of the AAAI Conference on Web and Social Media (ICWSM)*.

Mudde, C. 2004. The populist zeitgeist. *Government and opposition*, 39: 541–563.

Oliver, J. E.; and Rahn, W. M. 2016. Rise of the Trumpenvolk: Populism in the 2016 Election. *The ANNALS of the American Academy of Political and Social Science*, 667: 189–206.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Köpf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. arXiv:1912.01703.

Piazza, J. A. 2024. Populism and support for political violence in the United States: Assessing the role of grievances, distrust of political institutions, social change threat, and political illiberalism. *Political research quarterly*, 77: 152–166.

Prolific. 2025. Quickly Find Research Participants You Can Trust. https://www.prolific.com. Accessed 2025-01-15.

Qualtrics. 2025. Qualtrics XM — The Leading Experience Management Software. https://www.qualtrics.com. Accessed 2025-01-15.

Reimers, N.; and Gurevych, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the conference on empirical methods in natural language processing and the joint conference on natural language processing (EMNLP-IJCNLP)*, 3982–3992.

Rose, K. 2023. Gen Z's Social Media Dependency Is A Bridge, Not Barrier, For Advisors. Accessed: 2025-03-10.

Samory, M.; and Mitra, T. 2018. 'The Government Spies Using Our Webcams' The Language of Conspiracy Theories in Online Discussions. *Proceedings of the ACM Conference on Human-Computer Interaction (CSCW)*.

Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Starbird, K. 2017. Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on Twitter. In *Proceedings of the AAAI Conference on Web and Social Media (ICWSM)*.

Starbird, K.; DiResta, R.; and DeButts, M. 2023. Influence and improvisation: Participatory disinformation during the 2020 US election. *Social Media+ Society*, 9: 20563051231177943.

Stecula, D. A.; Kuru, O.; and Jamieson, K. H. 2020. How trust in experts and media use affect acceptance of common anti-vaccination claims. *Harvard Kennedy School Misinformation Review*, 1(1).

Steffen, E.; Mihaljevic, H.; Pustet, M.; Bischoff, N.; Varela, M. d. M. C.; Bayramoglu, Y.; and Oghalai, B. 2023. Codes, Patterns and Shapes of Contemporary Online Antisemitism and Conspiracy Narratives–an Annotation Guide and Labeled German-Language Dataset in the Context of COVID-19. In *Proceedings of the AAAI Conference on Web and Social Media (ICWSM)*.

The Knight Foundation. 2023. American Views 2022: Trust, Media, and Democracy. Accessed: 2025-05-14.

Tripodi, F. B. 2022. *The Propagandists' Playbook: How Conservative Elites Manipulate Search and Threaten Democracy*. Yale University Press.

Tripodi, F. B.; Garcia, L. C.; and Marwick, A. E. 2024. "Do your own research": affordance activation and disinformation spread. *Information, Communication & Society*, 27: 1212–1228.

Uscinski, J. E.; Enders, A. M.; Seelig, M. I.; Klofstad, C. A.; Funchion, J. R.; Everett, C.; Wuchty, S.; Premaratne, K.; and Murthi, M. N. 2021. American politics in two dimensions: Partisan and ideological identities versus anti-establishment orientations. *American Journal of Political Science (AJPS)*, 65: 877–895.

Wellman, M. L. 2024. "A friend who knows what they're talking about": Extending source credibility theory to analyze the wellness influencer industry on Instagram. *New Media & Society*, 26(12): 7020–7036.

Wike, R.; Fagan, M.; Huang, C.; Clancy, L.; and Lippert, J. 2025. Economic Inequality Seen as Major Challenge Around the World. Accessed 2025-04-23.

Zadrozny, B. 2021. Carol's Journey: What Facebook Knew About How It Radicalized Users. Accessed: 2025-04-23.

Zadrozny, B. 2025. More than 15,000 doctors sign letter urging Senate to reject RFK Jr. as health secretary. Accessed: 2025-04-03.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? [Yes]
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? [Yes]
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? [Yes]See Dataset, Human Annotation and Machine Annotation
   (e) Did you describe the limitations of your work? [Yes]See Limitations
   (f) Did you discuss any potential negative societal impacts of your work? [Yes]we discussed the potential negative impacts of our work. We concluded that this work has no negative impacts and instead has positive impacts by adding new scientific analyses of an important societal problem.
   (g) Did you discuss any potential misuse of your work? [Yes]
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? [Yes]We were worried that human annotators may be exposed to troubling content. We reviewed the annotation with IRB, collected informed consent, and put a trigger warning before each post. Annotators always had the option to not annotate a post if the content was upsetting to them.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? [N/A]
   (b) Have you provided justifications for all theoretical results? [N/A]
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? [N/A]
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? [N/A]
   (e) Did you address potential biases or limitations in your theoretical framework? [N/A]
   (f) Have you related your theoretical results to the existing literature in social science? [N/A]
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? [N/A]
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Partially yes. Please refer to the Empirical Evaluation section. The code is open-sourced.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]See Machine Annotation - Implementation Details
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]See Table 5
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? [Yes]
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? [No]We see limited to no cost and went to great lengths to get the highest quality annotations we could.
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes]See Human Annotation
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes]See Human Annotation and Findings
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? [N/A]
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? [N/A]
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? [Yes]See Table 1
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? [Yes]See Human Annotation

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes]See Human Annotation

(d) Did you discuss how data is stored, shared, and deidentified? [Yes]

Appendix

Visual Cues

TikTok videos contain rich visual information above and beyond what can be captured through transcribed audio. Here, we analyze whether there are differences in how content creators shoot their videos by topic. To do so we tag 738 videos according to their visual style. For example, as authenticity is an important component of AES, we might expect the conspiracy videos (the topic with the highest proportion of AES) to more often contain a person speaking directly to the camera. Alternatively, we may expect the opposite, as building authority can also be a component of AES.

In Table 7, we see that people are least likely to speak directly to the camera in the conspiracy content. This perhaps reflects that fact that the content creators in finance and wellness are more likely to be building a personal brand where centering the narrative around themselves is important. We also see that these categories are less likely to use embedded media than the conspiracy topic.

Next, we inspect how often the posts feature a person speaking directly to the camera in the conspiracy content that expresses AES and does not. Here, we see that in the conspiracy videos that express AES a person speaks directly to the camera 39% of the time and in the videos that do not express AES a person speaks directly to the camera 32% of the time. Thus, we see that overall conspiracy videos are less likely to feature a single person speaking to the camera than the other topics, however, AES content is slightly more likely to feature this point of view. This slightly supports the hypothesis that AES content creators choose visual cues that build authenticity, however, authenticity may be a less important factor for AES content than for content where personal brand building is important.

	Conspiracy	Finance	Wellness
Speaking directly to camera	.35	.57	.46
Speaking but not to camera	.34	.26	.31
Embedded media	.10	.05	.005
Text overlay with music	.06	.04	.07
Other	.15	.08	.145

Table 7: We inspect the ways the videos are shot for each of the three topics of conspiracy, finance, and wellness. Each entry is the proportion of content within this category that employs this visual style.

Information on Accounts Followed by Sock Puppets (For You Page Data)

We create 48 sock puppet accounts which interact with the For You Page. When the accounts are created they follow both news channels and lifestyle channels. The list of news channels is shown in Table 8 and the list of lifestyle channels is shown in Table 9.

author_name	author_id	Reliability	Bias
New York Times	nytimes	41.04	-8.07
NBC	nbcnews	42.80	-5.64
Washington Post	washingtonpost	38.83	-6.93
PBS News	pbsnews	43.32	-4.05
ABC	abcnews	44.80	-3.00
CBS	cbsnews	42.03	-2.72
NPR	npr	43.09	-4.17
BBC News	bbcnews	44.73	-1.35
Yahoo News	yahoonews	40.94	-5.63
USA Today	usatoday	40.86	-4.06

Table 8: Each account followed 6 news channels from the list above, where the channels were selected at random. Reliability scores and bias scores are extracted from the Media Bias Chart (Ad Fontes Media 2025). Reliability scores for articles and shows are on a scale of 0-64. Scores above 40 are generally good. Bias scores for articles and shows are on a scale of -42 to +42, with higher negative scores being more left, higher positive scores being more right, and scores closer to zero being minimally biased, equally balanced, or exhibiting a centrist bias. Typically, a publication would be considered centrist if the score is between -10 and +10, left-oriented if the score is -10 or less and right-leaning if the score is +10 or more. Thus, these accounts are all considered to be centrist.

Category	author_name	author_id
Sustainability	HomesteadDonegal	mirendarosenberg
	Ken Russell	kenforflorida
	Alaina Wood	thegarbagequeen
	Steeze365Daily	steeze365daily
	thesorrygirls	thesorrygirls
	ReLauren	relauren
	Reallaurinda	reallaurinda
	Ashley Diedenhofen	sciencebyashley
	TheNotoriousKIA	thenotoriouskia
	Phil Sustainability	philsustainability
	Brennan Kai	brennan.kai
Wellness	Jessamyn | The Underbelly Yoga	mynameisjessamyn
	Mari Llewellyn	marillewellyn
	Arielle Lorre	ariellelorre
	Dr. Will Cole	drwillcole
	Micheline Maalouf Therapist	micheline.maalouf
	Staci Tanouye, MD	dr.staci.t
	Andrew Huberman	hubermanlab
	Fiona	feelgoodwith_fi
	Steph Grasso, MS, RD	stephgrassodietitian
	Daniel	mrduku
DIY / Home improvement	Molly Miller	therenegadehome
	Bong Bain	wildheartshome
	Lilly	thefurnituredoctor
	Christine Higg	_forthehome
	Joanna Gaines	joannagaines
	Lone Fox	lonefoxhome
	Kylie Katich	kyliekatich
	CASSMAKESHOME | HOME & DIY	cassmakeshome
	Renovating Our Home	renovatingourhome
	Jay Munee DIY	jaymuneediy
	kelsey	kelseydarragh
	Abby	abby_roadhome
	Contractor Ken	contractorken
	ReallyVeryCrunchy	reallyverycrunchy
	THE FLIPPED PIECE	theflippedpiece
	Jeff Thorman	homerenovisiondiy
	Bro Builds	bro_builds
Tech	koharotv	koharotvreal
	Tyler Morgan	hitomidocameraroll
	Jimena con jota	soyjimenaconjota
	CHIP	chip_de
	Mark's Tech	markstech
	TheAsianJC	theasianjc
	Lucas VRTech	lucas_vrtech
	Marques Brownlee	mkbhd
	Unbox Therapy	unboxtherapyofficial
	Austin Evans	austintechtips
	iJustine	ijustine
	Kevin Stratvert	kevinstratvert
	Sara Dietschy	saradietschy
College Sports	Olivia Dunne	livvy
	Paige Beukers	paigebueckers
	Hanna & Haley Cavinder	cavindertwins
	Khoi Young	khoiyoung7
	Frederick Richards	frederickflips
	Shedeur Sanders	shedeursanders
	Angel Reese	angelreese10
	Caitlin Clark	caitlin.clark22
	Bronny James	bronny
	A.J. Henning	ajhenning
Hunting/fishing/outdoors	BlacktipH Fishing	blacktiph
	Ryan Izquierdo	ryanizfishing
	jetreef	jetreef
	RAWW Fishing	rawwfishingyt
	kickintheirbasstv	kickintheirbasstv
	Outdoors Weekly	outdoorsweekly
	Frederick Penney	frederick
	Becky Granola Girl	bonjourbecky
	Keith Paluso	thisiskeithpaluso
	Kween werK	kweenwerk

Table 9: Lifestyle influencers followed by the puppet accounts. Each account was assigned 2 topics and three followers from each topic.