Amira et al. *EPJ Data Science* (2025) 14:65
https://doi.org/10.1140/epjds/s13688-025-00582-6

EPJ Data Science
a SpringerOpen Journal

RESEARCH

Open Access

AGECovP: identifying ageism and analyzing COVID-19 discourse on older adults in YouTube

Ghenai Amira[1*], Nath Keshav[2] and Satsangi Aarat[3]

Handling Editor: Daniela Paolotti

*Correspondence:
aghenai@torontomu.ca
[1]Ted Rogers School of Management, Toronto Metropolitan University, Toronto, Canada
Full list of author information is available at the end of the article

Abstract

The COVID-19 pandemic significantly impacted older adults, generating widespread online discussions that revealed how this at-risk population was perceived. Understanding these portrayals is essential, as public discourse influences societal perceptions of aging and impacts policies and practices affecting older adults. Past research highlights that ageist stereotypes and attitudes frequently surface in public discussions, shaping the experiences of older individuals. The current study presents AGECovP, a comprehensive dataset featuring a diverse collection of YouTube videos, a leading social media platform. AGECovP is designed to provide researchers with meaningful insights into how older adults were portrayed during the pandemic and how topics such as conspiracy theories, misinformation, and the anti-vaccine movement were framed in relation to aging populations. In addition, the dataset includes a set of labeled comments indicating the presence of ageist content, enabling researchers to perform ageist detection and analyze ageism in online discourse. By providing a resource for examining both overt and subtle forms of ageism, AGECovP contributes to the development of tools and methodologies for addressing bias against older adults. This dataset fosters actionable insights into societal attitudes, enhancing the development of inclusive policies and interventions. Our data is available at: https://zenodo.org/records/15800324.

Keywords: Ageism Detection; Youtube; Older Adults; Covid-19

1 Introduction

Emerging in early 2020, The COVID-19 virus, identified as a highly contagious virus, has exhibited substantial morbidity and mortality rates. This resulted in profound challenges for the public health, healthcare systems, economies, communities, and families, as widely documented in global studies on the COVID-19 pandemic [1, 2]. This becomes especially crucial in times of health crises such as the COVID-19 pandemic, where older adults constitute a high-risk demographic.

Additionally, the Human Rights Council has highlighted that the pandemic has exposed deep-seated ageism and age discrimination across various domains. Older individuals have been unjustly held responsible for lockdowns and other measures that imposed re-

© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

strictions on personal freedoms [3]. Wister and Speechley [4] highlighted that the younger and working populations often perceive the COVID-19 pandemic as primarily affecting older adults, and they named it the "seniors problem". This perspective leads to the belief that younger individuals are less vulnerable and, even if infected, will experience milder symptoms compared to older age groups. Consequently, there is a sentiment that stringent economic and societal shutdowns may not have been necessary. In some instances, politicians have gone to the extent of suggesting that older individuals should consider sacrificing themselves for the greater good, encompassing both public health and economic considerations [5–7].

While the representation of older adults in traditional media have been well documented, examinations of such representations within social media discourse during a health crisis are still scarce. As YouTube is the most popular video-sharing website and the most visited one, this project will focus on the discussions around older adults on YouTube during the period of the COVID-19 pandemic.

In this work, we build AGECovP (Aging Perceptions in COVID-19 Pandemic), a curated dataset of YouTube videos from channels sharing information about COVID-19 and older adults. We aim to help researchers understand how older adults are generally represented in social media platforms during COVID-19 by analysing how individuals interacted with those videos from an aggregated (video) or individual (comments)[1] level. Our ultimate goal is to gain insights from the pandemic crisis, with a specific focus on analyzing the online discourse related to the population at higher risk, and to leverage these insights for future health crises. Additionally, AGECovP includes labeled comments for ageist content, enabling researchers to study and develop tools for automatic ageism detection in online discussions. AGECovP comprises the following information:

- Metadata of all 3782 videos that discussed content related to COVID-19 and older adults.
- Metadata of all 2243 channels that posted content related to COVID-19 and older adults.
- A total of 69,425 comment IDs of comments in all videos with basic metadata, sentiment, toxicity scores, and ageist content labels.
- A total of 4390 comment IDs with 'ageist' content labels, of which 1491 have been manually labeled.

AGECovP stands as a valuable asset for identifying ageist content and studying its prevalence and impact in online discussions related to older adults during the COVID-19 pandemic. By providing a curated dataset of YouTube comments and videos, it enables researchers to explore how ageist stereotypes and attitudes are expressed and propagated on social media. This dataset offers insights into the frequency and distribution of ageist language, helping to assess its broader societal impact and influence on public perceptions of aging. Additionally, it allows for the analysis of crucial themes such as misinformation, conspiracy thinking, and the anti-vaccine movement, highlighting how these topics intersect with ageism in online discourse. The contributions of this paper are as follows:

- Curate a large-scale dataset of videos with content about older adult during the period of the COVID-19 pandemic (https://zenodo.org/records/15800324).

[1] The comment text and real author names are not shared to protect the identity of the commenter. The actual comment text can berehydrated using YouTube Data API.

Table 1 Final set of search terms used. Each older adult-related search term was combined with each Covid-19-related search term for a total of 85 search term combinations

Older Adults Search Terms (N=17)
ageism, ageist, elderly, older, boomer, senior, aging, ageing, later life, age-related, retiree, retired, elders, geriatric, grandparent, grandmother, grandfather
Covid-19 Related Search Terms (N=5)
coronavirus, covid 19, covid-19, covid19, corona virus

- Perform exploratory analyses on the dataset to understand its key properties.
- Discuss potential uses for the dataset.

2 Data collection

We generate a list of terms used to describe the Covid-19 pandemic (N=5) and older adults (N=17), and then exhaustively searched for every combination of terms from the two lists (Table 1) using the Youtube Data API[2](Application Programming Interface), which allows programmatic access to YouTube video and comment metadata. The searches are executed with a daily total of 10,000 queries, adhering to the limitations imposed by the YouTube API. This resulted in 85 unique searches. For each combination, we collected all the results returned or *the first twelve pages* of results (*i.e.,* 600 results), whichever came first. By the twelfth page, results were not relevant to the topic. The videos were collected between the period of January 28, 2020, and August 21, 2022, and we specifically gathered videos in which the indicated regional language was English. We did not filter by uploader demographics or channel type; our aim was to capture all English-language COVID-19 and older adult-related videos on YouTube, regardless of whether they originated from individuals of any age group, news outlets, or commercial advertisers. Furthermore, no minimum or maximum duration filter was imposed. This resulted in a total of 6997 initial videos.

Out of the initially collected videos, some did not include the search keywords in either the title or the description. To ensure the videos were genuinely related to the topic, we applied regular expressions with the search keywords on both the title and description. Further, we eliminated the deleted videos from our dataset. This resulted in a total of 3353 videos.

Each YouTube video is linked to a list of "Suggested Videos". For each of the initial videos, we collected the list of suggested videos, which resulted in an additional 104,172 videos. To ensure that the suggested videos align with the topic of covid-19 and older adults, we filtered them based on keywords found in either the title or the description. Following this filtering process, the final list of suggested videos consisted of 1025 videos.

Merging the list of initially collected videos with the list from the suggested videos, the total number of videos reached 4378. Finally, due to the inaccuracies of the YouTube language detection tool, we employ a more reliable approach to filter out non-English videos. Specifically, we detected the language based on the titles and descriptions, using the Communalytic web application (https://communalytic.com) [8] able to detect English, French, German, and Russian. As a result, videos identified as having a non-English language are excluded from the analysis. The final number of videos included in the AGECovP dataset

[2] https://developers.google.com/youtube/v3.

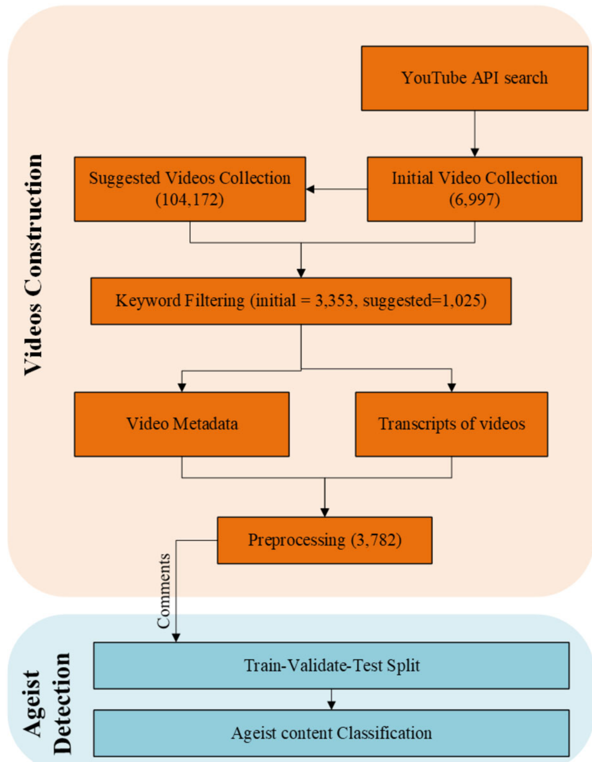


Figure 1 Overview of AGECovP Collection

after the adopted cleaning steps was 3782. All engagement metrics (e.g., views, likes, comments, subscribers) reflect a single snapshot taken at the time of data collection (January 2020–August 2022). These values are not updated, so the dataset captures engagement as it existed during the collection period. Specifically, all content in the AGECovP dataset originates from user-generated videos and comments posted on YouTube; the research team only curated and annotated this online content for research purposes. Figure 1 is a fowchart that summarizes the dataset construction proposed in this paper.

Video metadata We collected the metadata for every video retrieved by our search queries, and for each video, we also recorded metadata about the corresponding channel that published it. In this context, channels refer to the YouTube accounts responsible for uploading the videos included in our dataset. The metadata includes the title, description, category, country, tags, number of likes, number of views, number of comments, duration, and published date. The "category" field in our dataset is assigned automatically by the YouTube platform and reflects standardized categories provided by the YouTube Data API. To get more information about the videos' content, we collect the transcripts or subtitles provided by YouTube. For each video, we use the Detoxify API[3](an open-source machine learning tool for detecting toxic language in text) to perform scoring for toxicity, profanity, and insult on the fields of the video title aggregated with the description.

[3]https://github.com/unitaryai/detoxify.

Additionally, we applied two distinct lexicon and rule-based libraries, namely TextBlob[4] and VADER,[5] to assess the sentiment expressed in the videos looking at the title as well as the description fields. These libraries have undergone empirical validation and have been widely employed by researchers [9–12], demonstrating their effectiveness across various domains.

We further collect information about every channel that posted a video in our dataset. A channel on YouTube is a user's or organization's personal space where they can upload and organize videos, interact with viewers, and build a subscriber base. Each channel has a unique profile and may represent an individual, group, media outlet, or institution. The metadata includes the title, description, country, age, number of subscriptions, and a list of categories describing the channel content.

Finally, we collect all the comments received for each video. For each comment, we store the date, number of likes, number of replies, and parent id (in case the comment is not first level comment). We filter out comments where the language detected from the text is not English. Similar to each video, we use the Detoxify API to perform scoring for toxicity, profanity, and insult and we adopt TextBlob and VADER to assess the sentiment. Additionally, we provide a label for each comment indicating whether it contains ageist content or not. The details of the labeling process, including how the comments were manually coded for ageism, are explained in Sect. 5. Table 2 summarizes all the fields available in the AGECovP dataset.

3 Dataset overview

Most of the videos are less than 5 minutes long with an average of 2.7 minutes (minimum less than 1 min and a maximum of 715 minutes). The most dominant category is "News and Politics", followed by "People and Blogs" and 49.6% of the videos were originated from the USA. Figure 2(a) displays the daily distribution of videos in the AGECovP dataset, with a 7-day moving average overlaid to highlight broader trends. While a prominent peak is visible in March 2020—corresponding to the World Health Organization's declaration of COVID-19 as a global pandemic—several other peaks are also observed throughout the study period. Notably, increased video activity in January 2021 may be associated with the FDA approval and rollout of COVID-19 vaccines. The moving average line facilitates interpretation of these patterns by smoothing out short-term fluctuations and emphasizing multiple periods of heightened video activity, reflecting responses to major pandemic-related events.

Looking at the engagement around AGECovP dataset videos, Fig. 3 presents the cumulative distribution functions (CDFs) of views, likes, and comments per video, with all values plotted on a $\log_{10}$ scale. The distributions reveal that most videos receive moderate engagement, with approximately 80% of videos accumulating between 100 and 10,000 views, the majority of likes ranging from 10 to 1000, and most comments falling between 10 and 100. Only a small fraction of videos achieve very high engagement, reaching millions of views, tens of thousands of likes, or thousands of comments. This illustrates that while viral content exists, the vast majority of videos maintain lower levels of user interaction. Further, Fig. 4 shows a highly skewed pattern of subscribers per channel, where

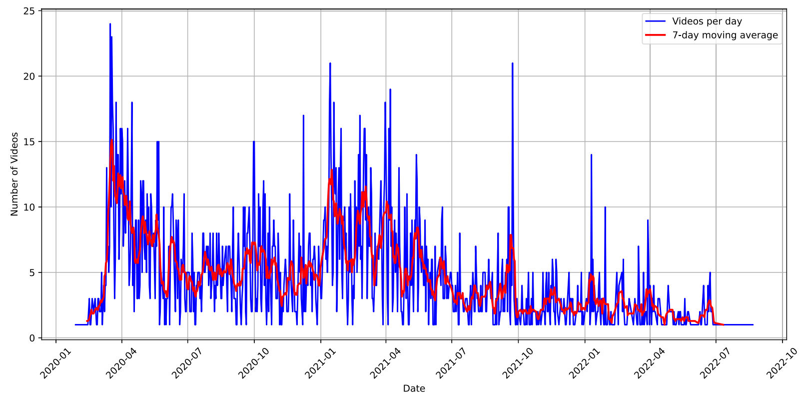
[4] https://textblob.readthedocs.io/en/dev/.

[5] https://github.com/cjhutto/vaderSentiment.

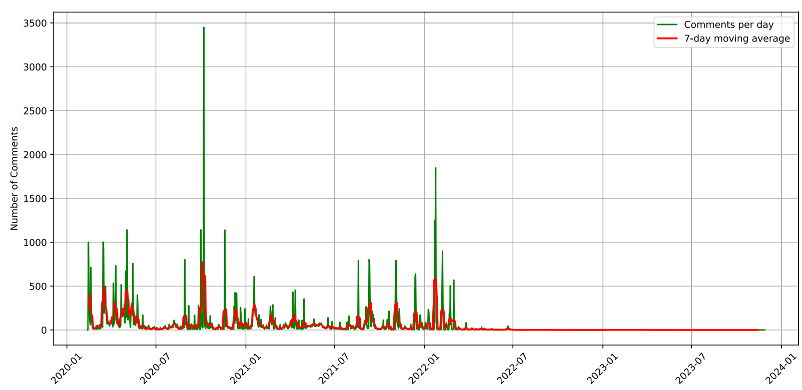
Table 2 Fields and descriptions in the AGECovP dataset

Field	Description
Videos	
Video_id	ID that YouTube uses to uniquely identify a video
Channel_id	ID that YouTube uses to uniquely identify a channel
Duration	Duration of the video (minutes)
Published_date	Date the video was published
Title	Title of the video
Description	Description of the video provided by content creator
Category	Content category provided by Youtube
Country	The country with which the content creator is associated with
Tags	Tags provided by content creator
Caption	Transcript of the video from the youtube-transcript-api
Views	Number of views
Likes	Number of likes
Comments	Number of comments
Toxicity	Toxic score from Detoxify API
Profanity	Profanity score from Detoxify API
Insult	Insult score from Detoxify API
VADER_SenV	Sentiment score from VADER API
TextBlob_SenV	Sentiment score from TextBlob API
Channels	
Channel_id	ID that YouTube uses to uniquely identify a channel
Title	Title of the channel
Description	Description of the channel
Country	The country with which the channel is associated with
Age	Number of days since the channel was created
TopicCategories	List of Wikipedia URLs that describe the channel's content
Subscriptions	Number of subscriptions for the channel
Comments	
Comment_id	Unique identifier automatically assigned by YouTube to every comment posted for a video
Parent_id	ID of the parent comment in case the comment is a *reply* to another comment.
Video_id	ID of the video comment is on
Anon_id	anonymised channel's id
LikeCount	Number of likes on comment
PublishedAt	Publish time of comment
ReplyCount	Number of replies to comment
Toxicity	Toxic score from Detoxify API
Profanity	Profanity score from Detoxify API
Insult	Insult score from Detoxify API
VADER_SenC	Sentiment score from VADER API
TextBlob_SenC	Sentiment score from TextBlob API
Ageist Label	
Comment_id	Unique identifier automatically assigned by YouTube to every comment posted for a video
Ageist_label	1 if the comment contains ageist content 0 otherwise

most channels have relatively few subscribers, with a peak around 1000 subscribers. As the subscriber count increases, the number of channels decreases significantly, with very few channels reaching over a million subscribers. The AGECovP dataset represents a broad spectrum of channels, including both fringe creators with fewer subscribers and larger, more influential channels. However, most videos come from channels with smaller subscriber counts, which suggests that the dataset leans more toward creators who may not be as heavily promoted by YouTube's algorithm. Despite this, the presence of videos with high engagement demonstrates that certain content has the potential to reach broader audiences, even if the channel itself is not among the largest on the platform. This mixture provides a balanced view of both lesser-known and more popular creators.



(a) Distribution of Videos



(b) Distribution of Comments

Figure 2 Daily volume of videos and comments in the AGECovP dataset, with 7-day moving average

YouTube is designed as a platform for sharing user-generated content, but a considerable number of videos on the platform are reproductions sourced from elsewhere, including movies, TV shows, and other professional video websites, rather than being original creations by ordinary internet users. To uncover the nature of videos in our dataset, we categorized them as either a UGC (User Generated Content) or UCC (User Copied Content) video where UCC is a snippet from external sources by professional producers and UGC is a video created for user-captured content [13]. Specifically, we conducted manual assessments of each video's content by examining descriptions or viewing the videos themselves to classify them as UGC or UCC. To ensure consistency, two authors independently coded an initial sample of 10% of the videos. Any disagreements were discussed and resolved collaboratively to establish a shared understanding of the coding criteria. After reaching an agreement, one author proceeded to label the remaining videos based on the agreed-upon guidelines. Results of this labeling process showed that less than 7% of the videos were UCC and not UGC. This indicates that the dataset is dominated by videos originally aired on television by news sources, among others, and very few videos were originally created by YouTube users using their digital cameras and webcams.

After the data collection phase, here we summarize basic statistics about the channels that uploaded the AGECovP videos. The dataset contained user accounts with an average

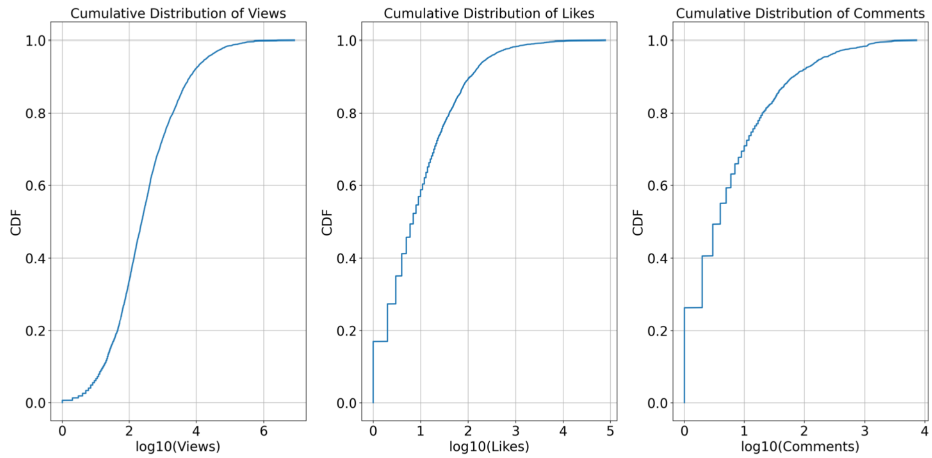


Figure 3 Cumulative distribution functions (CDFs) of video engagement metrics on a $\log_{10}$ scale, showing the distribution of views, likes, and comments for all videos in the dataset

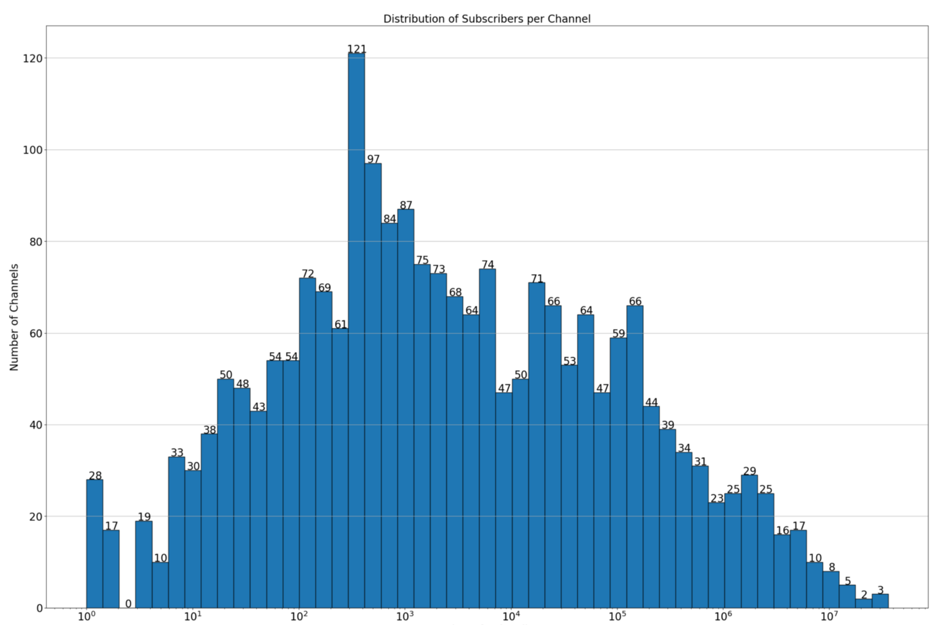


Figure 4 Distribution of Subscribers per Channel

account age of 2931 days, calculated as the number of days since each account's creation date as provided by the YouTube API (ranging from a minimum of 65 days to a maximum of 6163 days). The average number of subscribers for channels in the dataset was 443,891, with a maximum of 36,100,000. Additionally, YouTube API offers insights into the topics associated with the channel content. Analyzing users who posted content within the AGE-CovP dataset, the prevalent categories/topics were Society (56.4%), followed by Lifestyle (sociology) (13.6%), Knowledge (12.1%), and then Politics (4.1%) and Health (3.7%).

Regarding the comments, our data collection spanned from February 12, 2020, to May 20, 2023. We made sure to continuously retrieve the latest comments associated with the initially gathered videos where we crawled new comments every 7 days. Figure 2(b) shows the distribution of the comments by day where "Comments" are top-level comments to videos and "Replies" are replies to top-level comments. The comments encompassed 47,706 (68.7%) direct responses to the videos and 21,719 (31.3%) replies to existing comments. These figures underscore a significant level of interaction among users within the dataset. On average, the comments received 6 likes, with a maximum of 10,976.

In the following sections, we provide an analysis of the AGECovP dataset to gain useful insights that will assist in demonstrating the opportunities opened by this new dataset.

4 Exploratory data analysis

In this section, we conduct a detailed investigation of the AGECovP dataset with a particular focus on uncovering patterns that may indicate ageist content. The primary objective of this analysis is to explore underlying trends in sentiment, toxicity, and topics to better understand how older adults are portrayed in online discussions during the COVID-19 pandemic. By examining the tones, narratives, and common themes in both video content and comments, we aim to identify areas where ageist stereotypes or attitudes may be present. For the upcoming analysis, we conducted topic, sentiment, and toxicity analysis using Communalytic platform (a research tool for studying online discourse which operates under a freemium license) [8].

4.1 Topic analysis

The goal of the topic analysis is to uncover key themes within the dataset that may reveal how older adults were discussed during the COVID-19 pandemic. By identifying recurring topics in both videos and comments, we aim to explore potential ageist content, societal perceptions, and concerns related to older adults. This allows us to understand the broader narratives surrounding aging and how these narratives may contribute to or combat ageism in public discourse.

For topic modeling, we first generate semantic sentence embeddings for each text entry (video description or comment) using pre-trained sentence-transformer models from Hugging Face [14], specifically the Sentence-BERT (SBERT) framework. This approach transforms each sentence into a high-dimensional vector that captures contextual and semantic meaning, enabling more nuanced similarity detection compared to traditional bag-of-words models.

To cluster and visualize these embeddings, we utilize the Nomic Atlas platform,[6] which projects the high-dimensional sentence vectors into a two-dimensional interactive map using dimensionality reduction techniques (such as UMAP or t-SNE). Atlas then applies unsupervised clustering algorithms (e.g., HDBSCAN) to group similar texts together, automatically generating initial topic labels based on the most representative keywords within each cluster. The initial clustering and topic labeling is thus software-generated, based on semantic similarities and the most representative keywords in each cluster.

After the automatic clustering, two members of the research team independently reviewed the clusters and their top keywords. We examined representative comment samples from each cluster to ensure that the software-generated labels and groupings reflected

[6]https://atlas.nomic.ai/.

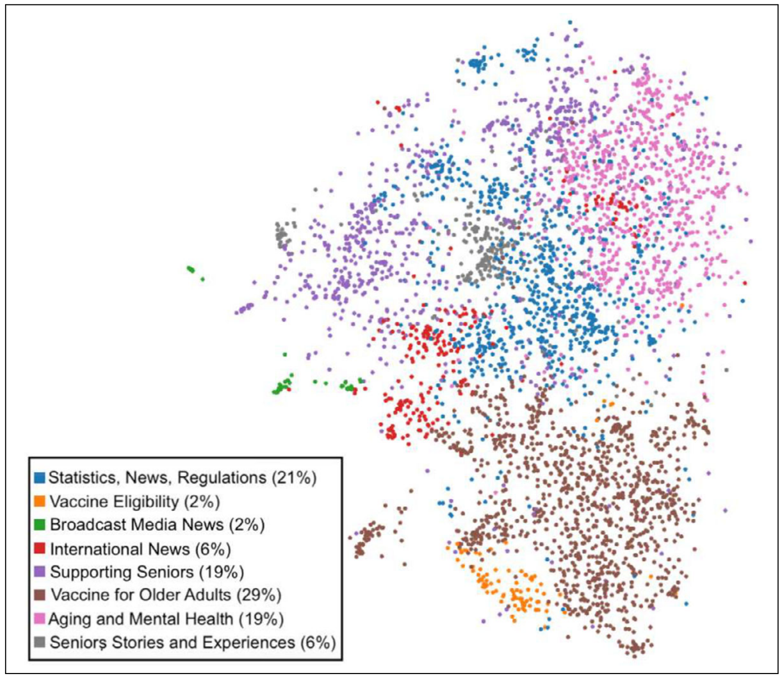


Figure 5 Topics Representation of AGECovP Videos

meaningful and interpretable topics. In cases of ambiguity or disagreement, the two researchers discussed and reached consensus on the final topic label and its description. This expert-in-the-loop approach ensured that the resulting topics are both data-driven and contextually relevant for the study of ageist content in COVID-19 YouTube comments. This process is performed separately for both video descriptions and comments, resulting in the identification of eight key topics in each modality.

This hybrid approach of automated topic discovery with expert manual validation allows us to balance scalability with interpretability and ensures that the identified topics are meaningful for subsequent analysis.

Topic analysis on videos Figure 5 shows the sentence embeddings of the videos' descriptions, visually presented using Nomic Atlas. We provide detailed descriptions of the most important topics as follows.

The most dominating topic, "Vaccine for Older Adults" (1127 videos, shown in *brown* in Fig. 5), focuses on the importance of vaccinating high-risk populations, particularly older adults. It discusses vaccination clinics tailored for older adults and largely conveys a positive sentiment, emphasizing the crucial role of vaccination in overcoming the pandemic. The next prominent topic, "Statistics, News, Regulations" (809 videos, shown in *blue* in Fig. 5), covers pandemic-related news, including COVID-19 statistics, policies, and regulations affecting older adults. This theme generally carries a positive sentiment. The third topic, "Supporting Seniors" (731 videos, shown in *purple* in Fig. 5, focuses on initiatives to assist older adults, such as volunteer work, food delivery, and telehealth training. The sentiment is largely positive, reflecting support for at risk populations during the pandemic. The final topic, "Aging and Mental Health" (736 videos, shown in *pink* in Fig. 5, explores

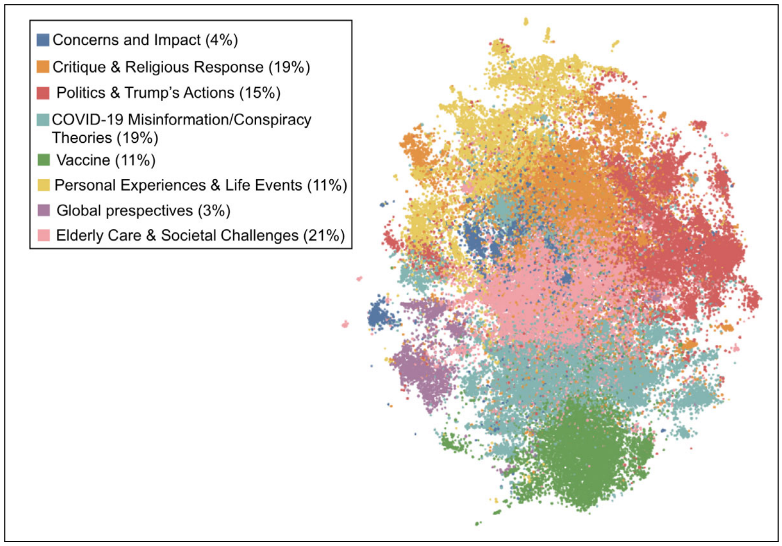


Figure 6 Topics Representation of AGECovP Comments

the effects of social distancing on older adults, COVID-19's impact on older adults, and regulations in care homes. Most content focuses on helping the elderly cope with these challenges, promoting community kindness.

Overall, the topics focus on pandemic-related news, regulations, support initiatives, and mental health, with a generally positive sentiment emphasizing the support for older adults during the crisis. However, much of the content is driven by news outlets rather than user-generated material, potentially reflecting editorial biases rather than authentic public views on older adults. In the next section, we explore how user-generated content portrays this demographic.

Topic analysis on comments In the comments, the most dominant topic, "Elderly Health and Societal Challenges" (14,468 comments, shown in *pink* in Fig. 6), centers on concerns about the elderly, health issues, and societal challenges. Many users expressed frustration with regulations affecting older adults, often leading to feelings of isolation. Notably, 44% of the comments carried negative sentiment, with claims that COVID-19 was created to target the elderly or blaming them for strict pandemic regulations. One user commented (rephrased to preserve user's privacy), "*Have they not lived a long life already? Our concern should be the younger generation*." These discussions reveal significant negativity towards older adults, indicating the importance of further investigation into ageist content.

The next most prevalent topic, "COVID-19 Misinformation/Conspiracy Theories" (13,415 comments, presented in *turquoise* in Fig. 6), is related to misinformation, conspiracy theories, and skepticism surrounding COVID-19. The words mentioned, such as "Vaccine injuries," "Virus Hoax," "Pfizer's Black Magic Genetics," and "Jabbed rats," suggest a narrative that questions the safety and efficacy of vaccines, promotes conspiracy theories, and may involve a distrust in scientific information. Further, the common reference to the terms "Commiefornia" and "CCP" indicates a potential political or ideological dimension to the discourse. An example (rephrased) of a comment from this topic is: "*Playing with emotions; we acknowledge the loss, but argue that deaths weren't due to the alleged*

Table 3 Sentiment Analysis by VADER and Textblob libraries where Negative Sentiment is a score between [-1..-0.05], Neutral Sentiment is [-0.05..0.05], and Positive Sentiment is [0.05..1]

Sentiment	Negative	Neutral	Positive
Videos			
VADER	905 (21%)	860 (20%)	2486 (59%)
TextBlob	251 (6%)	934 (22%)	3066 (72%)
Comments			
VADER	26,804 (39%)	15,276 (22%)	27,345 (39%)
TextBlob	15,052 (22%)	27,415 (39%)	26,958 (39%)

hoax virus. They resulted from other factors or old age, possibly worsened by this year's common cold strain, also called coronavirus."

The topic "Critique and Religious Response" (11,883 comments, shown in *orange* in Fig. 6) focuses on criticism of COVID-19 regulations, with a religious dimension where individuals turn to their faith during challenging times. Comments also address racism, particularly towards Asians and the Chinese community, while others reflect on using religion for strength in facing the pandemic. The topic "Politics & Trump's Actions" (10,365 comments, shown in *black* in Fig. 6) centers on political discussions, including critiques of news coverage, comments on figures like Trump, and debates on issues like abortion and women's rights. It explores leadership evaluations and current events, with a particular focus on the Trump administration. Notably, 43% of the comments in this topic carried a negative sentiment. One example comment states, "*Trump neglected his responsibilities by ignoring intelligence reports about the pandemic, dismissing it as a DNC hoax, and failing to act during a critical six-week period.*" Moving away from politics, the "Vaccine" topic (7334 comments, shown in *green* in Fig. 6) discusses vaccines, immunity, side effects, government policies, and the vaccination process.

Although the videos were curated using keywords related to COVID-19 and older adults, the comments exhibit a wide range of topics, including misinformation, conspiracy theories, political critiques, vaccines, and personal experiences. While not all comments focus directly on older adults, the discussions often intersect with issues affecting this demographic, shaping perceptions of their role during the pandemic. These varied topics provide insight into how older adults were viewed, sometimes sympathetically, but often in a negative light, particularly in discussions related to regulations and public health measures.

4.2 Sentiment analysis

We used two lexicon-based tools, TextBlob [15] and VADER [16], to obtain sentiment scores and compared their performance to determine which provided the most accurate results for our dataset. The following section presents the sentiment analysis results for both videos and comments.

Videos sentiment analysis We apply sentiment analysis to both the titles and descriptions of videos. Table 3 shows the results, while Fig. 7 (a) compares TextBlob and VADER outcomes. TextBlob shows 72% of videos having a positive content, 22% are neutral, and 6% negative. VADER shows 59% positive, 20% neutral, and 21% negative. Despite slight differences, both tools indicate that most videos convey positive sentiments about COVID-19 and older adults, with only a small portion showing neutral or negative content.

To further examine the relationship between the two sentiment analysis tools, we include a scatter plot comparing VADER and TextBlob sentiment scores for all videos (see Fig. 7 (b)). The Pearson correlation coefficient is 0.30 ($p < 0.00001$), and the Spearman correlation is 0.33 ($p < 0.00001$), indicating a moderate but not strong association between the two methods. This visualization highlights the extent of agreement and disagreement between VADER and TextBlob at the individual video level.

We additionally analyze cases where VADER and TextBlob disagree (413 videos with positive VADER scores but negative TextBlob, and 72 vice versa). VADER and TextBlob agree on 60% of the videos, showing a *fair* level of agreement (Cohen's kappa = 0.213). One of the authors manually investigated a random sample of 100 (50 instances from where Vader assigned positive scores and Textblob assigned negative scores and 50 instances vise versa) by reading the descriptions, and listening to the videos when in doubt. The results showed that VADER was more accurate in 78% of cases, though its performance degraded with longer, more nuanced texts. This highlights the limitations of lexicon-based tools in capturing complex sentiments, especially when evaluating content related to older adults and COVID-19.

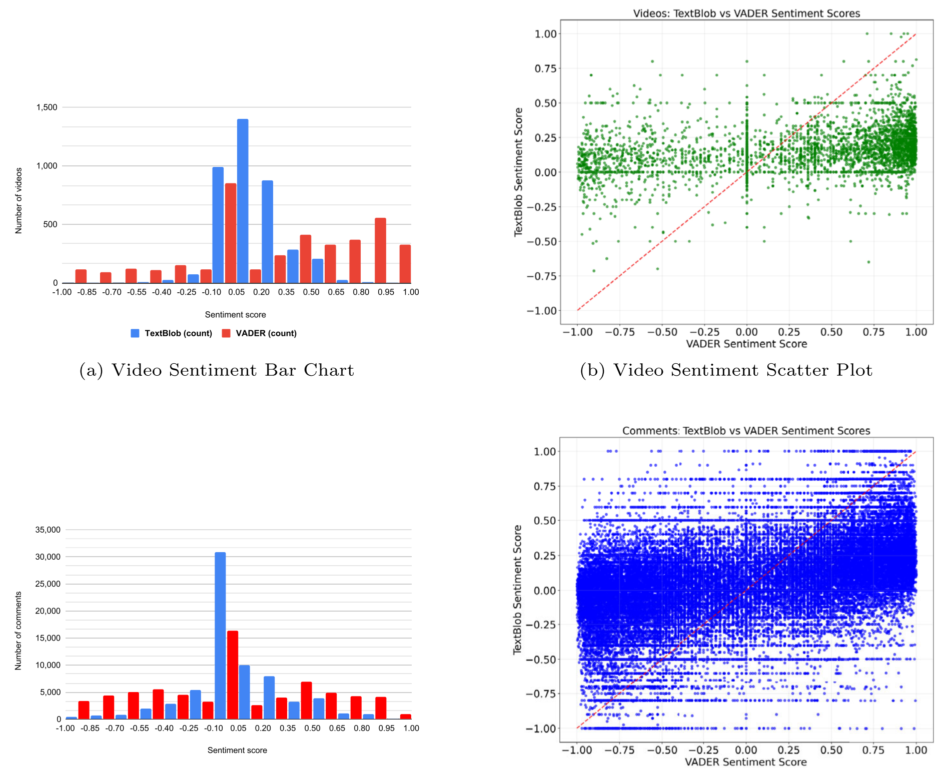
The sentiment analysis shows a predominantly positive sentiment across the videos (59% in VADER, 72% in TextBlob), which may seem unexpected given the topic. This could be due to the dominance of positive themes in the extracted topics, as well as the neutral language used in videos, many of which are from news outlets rather than user-generated content (UGC accounts for less than 7%).

Comments sentiment analysis Similarly, we capture the sentiment in the videos' comments. Table 3 shows the results and Fig. 7 (c) shows the distribution of sentiment scores computed by both libraries. According to TextBlob, 39% of the comments have neutral sentiment, followed by 39% having positive sentiment, and the remaining 22% are negative. In VADER, positive and negative sentiments are evenly distributed, each accounting for 39% of the data, while 22% of the comments are classified as neutral. While there exists a disparity in the output of the two sentiment analysis libraries, a noticeable trend emerges, revealing a higher proportion of negative content in the comments compared to the sentiment expressed in the videos.

To further examine the relationship between VADER and TextBlob sentiment scores for comments, we include a scatter plot comparing the two methods (see Fig. 7 (d)). The Pearson correlation coefficient is 0.45 ($p < 0.00001$), and the Spearman correlation coefficient is 0.45 ($p < 0.00001$), indicating a moderate positive association between the two sentiment tools. This visualization highlights both the agreement and divergence at the individual comment level.

VADER and TextBlob agree on the categorization of 37,479 comments (54.34%) with a fair level of agreement (Cohen's kappa = 0.325 [17]). To further explore the discrepancies, one of the authors manually examined comments with conflicting polarity scores: 2720 cases where VADER assigns a positive score and TextBlob assigns a negative score, and 6831 cases where the opposite occurs.

From these, we select 100 random comments for manual investigation. The results show that VADER has a higher accuracy rate, correctly identifying sentiment in 61% of cases compared to TextBlob. However, the two tools perform differently depending on the content. TextBlob is more sensitive to profanity, while VADER excels in handling emoticons



(a) Video Sentiment Bar Chart

(b) Video Sentiment Scatter Plot

(c) Comment Sentiment Bar Chart

(d) Comment Sentiment Scatter Plot

Figure 7 Comparison of sentiment scores for AGECovP dataset videos and comments: Bar charts (a, c) and scatter plots comparing VADER and TextBlob sentiment scores (b, d) for videos and comments

[18], sarcasm, negation, questioning, slang, and capitalization. Despite VADER's strengths in these areas, TextBlob ultimately provides better overall performance for analyzing the AGECovP comments.

Overall, the results reveal a diverse range of opinions and emotions in the comments related to older adults during the COVID-19 pandemic. Notably, the higher proportion of negative sentiment in the comments compared to the videos indicates a more critical or unfavorable stance toward older adults in public discussions. This suggests that while video content may present older adults in a generally supportive or neutral light, user-generated comments tend to reflect more negative or biased views, often involving frustrations or blame directed toward this demographic. These findings highlight the potential presence of ageist content, as negative sentiments toward older adults can contribute to stereotypes or harmful narratives, further emphasizing the need for targeted detection and analysis of ageism in online discourse.

It is important to note that the fair level of agreement between TextBlob and VADER in both the videos and comments highlight both the ambiguity of sentiment in our dataset and the limitations of general-purpose sentiment analysis tools when applied to complex social topics. This finding emphasizes the importance of using multiple methods and considering manual or domain-adapted approaches for a more nuanced understanding of sentiment in online discourse.

Table 4 Sample Comments with Corresponding Toxicity Scores

Comment	Score
"And the rest just die because they are seniors God I am sick of this covid nonsense"	0.96
"Are you crazy lol 1 in every 100 seniors die from covid Do you realize if that was the case we would have over a million deaths just in seniors citizens these fools are ridiculous and crazy to even make a statement like that and this is why people dont trust or watch CNN because they are all liars"	0.99
"then the unsubsidized countries mutated the virus and sent it right back to the boomers Fuck em all"	0.96
"happyfrown Yeah anyone who doesnt support a kid president or kids consenting to anything are ageist Pedophiles frothing at the mouth right now"	0.82
"I want to move Thanksgiving to Friday so the elderly can have a good dinner before they die Joe"	0.03
"So 1 of seniors huh We should shut down the country and live in fear if thats the case"	0.02

4.3 Toxicity analysis

We collect toxicity scores from the Detoxify API [19] for the AGECovP dataset, reporting on toxicity, profanity, and insult values. In this study, "toxicity" refers to the degree of rude, disrespectful, or harmful language present in a comment, as detected by the Detoxify API. This encompasses not only biased or stigmatizing language but also insults, hostility, and other forms of abusive discourse. In the AGECovP dataset videos, toxic language is minimal, with over 90% of them scoring less than or equal to 0.3 for toxicity, profanity, and insult. This is likely due to the fact that most videos come from news media and informative outlets, which typically employ moderated, neutral language.

Looking at the comments, however, we observe a higher prevalence of toxic content. Specifically, 83% of comments have toxicity scores below 0.3, 4% have scores between 0.3 and 0.5, and 13% score between 0.5 and 1. This shows that toxic language is more common in user-generated comments compared to the videos. Table 4 presents a sample of the most toxic content, some of which includes harmful sentiments directed towards older adults. These comments often trivialize the health of older adults, spread conspiracy theories, or promote misinformation by questioning the legitimacy of the pandemic and government actions.

Despite the utility of toxicity analysis, it does not fully capture the nuances of ageism. Many comments that perpetuate ageist ideas or stereotypes may not score high on toxicity but still contribute to harmful discourse. For example, comments that frame older adults as inherently weak or suggest that their deaths are inevitable due to age may not register as toxic but clearly convey ageist attitudes (See the last two examples in Table 4). This limitation highlights the need for additional labeling that focuses specifically on identifying ageist content.

In the next section, we introduce a dedicated labeling process aimed at detecting ageism, allowing us to more effectively capture and analyze such content within the AGECovP dataset.

5 Ageist content detection

In this section, we aim to identify ageist content in user-generated comments to better understand how older adults were portrayed during the COVID-19 pandemic. However, the challenge of identifying ageist content in such comments is not fully addressed by traditional sentiment analysis or toxicity detection tools. These automated methods often miss subtle or satirical ageist remarks that may appear neutral or even positive in sentiment.

For example, a comment like "*1 in 100 seniors dying of covid within the last year and a half Why does this NOT SOUND INCREDIBLE*" was given a neutral sentiment score of 0.1333 by TextBlob and an extremely low toxicity score of 0.0011, despite clearly containing sarcastic and ageist language. Such cases illustrate the shortcomings of relying solely on sentiment or toxicity metrics, as they fail to capture the discriminatory undertones or dehumanizing content directed at older adults.

5.1 Manual labeling guidelines

We implement a manual labeling process to accurately identify ageist content. We begin by employing a two-step filtering approach to extract relevant comments. First, we apply a regex-based filter to identify comments containing explicit age-related keywords [7] [20–22]. The keywords used to extract ageist content cover a broad range of terms, from neutral descriptors like "older adults" and "senior citizens" to more clearly derogatory or ageist phrases like "useless old people," "fossil," and "washed up." This diversity in the terms is crucial for detecting ageist content, as ageism can manifest in various forms—both subtle and explicit.

This initial filter yielded 3854 comments from the dataset. Recognizing that relevant comments may not always include these specific keywords, we then apply an advanced embedding-based semantic filter to the remaining 37,293 comments that did not match the regex. For this, we use the pre-trained Sentence-BERT (SBERT) model, paraphrase-MiniLM-L6-v2 [14], which allows us to capture the semantic meaning of comments. This additional filtering identified 536 additional comments that are contextually relevant to ageism or aging. In total, 4390 comments are selected for further analysis through manual labeling.

During the labeling process, one of the authors – an expert in older adults research and ageism concepts – develops comprehensive labeling guidelines. These guidelines define ageism as prejudice, discrimination, or stereotyping based on a person's age, focusing on derogatory comments about older, age-based generalizations (e.g., "All elderly people are senile") [23]. Each comment is assigned one of two labels: *TRUE* for clearly ageist content expressing prejudice or stereotypes, *FALSE* for non-ageist content reflecting concern or neutrality, or for off-topic or unrelated content. Regarding the types of ageism, our labeling guidelines are designed to capture a wide spectrum of ageist content, encompassing both personal and institutional ageism, and including explicit (overt) as well as implicit (subtle, coded) forms.

First, an initial sample of 500 comments is selected to ensure the validity of the coding guidelines and improve the reliability of the labels. Two authors independently code the same set of comments using the developed guidelines. Afterward, they meet to discuss any disagreements, and Cohen's Kappa is calculated to measure inter-rater reliability, resulting

[7] Full keyword list: elderly, older adults, seniors, senior citizens, aging population, old people, geriatric, retired individuals, the aged, advanced in age, golden years, over the hill, getting old, senior living, older generation, silver-haired, aging parents, retired workers, nursing home, assisted living, long-term care, home care, retirement home, elder care, dementia care, hospice care, senior care, memory care, senior health, age-related diseases, Alzheimer's disease, dementia, cognitive decline, frailty, osteoporosis, mobility issues, too old to, set in their ways, old-fashioned, senile, out of touch, slow learners, stubborn old people, past their prime, too slow, cranky old people, can't keep up, losing it with age, inflexible, can't teach an old dog new tricks, ageist, ageism, age discrimination, age bias, anti-aging, burden on society, useless old people, aging out, old and irrelevant, aging workforce, too old for this, dealing with the elderly, older people are a burden, ageism in the workplace, retirement age, pensioner, social security, retirement fund, retirement planning, retired life, senior discounts, pension plan, early retirement, aging into retirement, post-retirement, financial planning for seniors, respect your elders, generational gap, fossil, old fart, old geezer, old bag, dinosaur, dead weight, washed up, past their prime.

in a score of 0.79, indicating "substantial agreement" according to the commonly used scale [24]. For cases where disagreements remain after the discussions, a third coder reviews the comments, and a majority vote is used to assign the final label. After that, one coder additionally labels another 568 comments with a total of 1068 comments from the overall dataset.

5.2 Automated ageist content detection

After manually annotating 1068 comments, we identified a significant class imbalance, with only 23% labeled as ageist and 77% as non-ageist. To address this, we fine-tuned a RoBERTa-base transformer model to perform binary classification using the annotated comments. The labeled dataset was split into training (70%), validation (15%), and test (15%) sets, with comment text serving as the input and the ageist/non-ageist label as the output. Model performance was evaluated using accuracy, precision, recall, and F1 score for both classes. The final model achieved an accuracy of 0.75, with an F1 score of 0.64 for the ageist class and 0.8 for the non-ageist class. The trained model was then applied to the remaining unlabeled comments, assigning pseudo-labels of ageist or non-ageist to expand the dataset. To ensure the quality of the newly labeled data and increase the number of confirmed ageist examples, we randomly selected 423 comments that had received the "ageist" pseudo-label from the model and manually annotated them using the same coding guidelines, with each comment independently reviewed by two human annotators. This iterative approach yielded a final dataset of 4390 comments, with 810 (19%) labeled as ageist. The reliability of the manual annotations was assessed using Fleiss's Kappa [25], resulting in a value of 0.5134, which is interpreted as moderate agreement according to the guidelines of Landis [24]. This level of agreement reflects both the inherent complexity of identifying nuanced ageist content in online discussions and the current limitations of automated methods, while supporting the consistency and value of our human annotation guidelines.

5.3 Ageist content exploratory data analysis

5.3.1 Descriptive analysis of ageist comments

To better understand the nature and characteristics of ageist content in the AGECovP dataset, we performed an exploratory data analysis focusing on all comments labeled as ageist.

Toxicity and Sentiment Characteristics: Ageist comments tend to be more toxic and negative in sentiment than the general comment pool. The average toxicity score for ageist comments is 0.17 (SD = 0.29), but the distribution is highly skewed: while the median toxicity is only 0.008, a subset of comments exhibits extremely high toxicity (up to 0.997). Similarly, profanity and insult scores are generally low in the median but reach high values in the most toxic examples.

Sentiment analysis using VADER reveals that ageist comments are on average negative (median = -0.18). However, the sentiment distribution is broad, with a minority of comments expressing positive or neutral sentiment. In contrast, TextBlob assigns a slight positive mean sentiment (mean = 0.07, median = 0.06), but the distribution again spans the entire range from strongly negative to strongly positive. These differences reflect the nuances and complexities of detecting ageist content; ageist remarks may occur in comments that are not overtly toxic or negative.

Engagement and User Behavior: Most ageist comments receive limited engagement. However, a few ageist comments achieve significant visibility, garnering up to 204 likes and multiple replies. Analysis of the user base shows that most users post only a single ageist comment, though a small number of users are responsible for multiple such comments. Similarly, certain videos attract a disproportionately high number of ageist comments, with the top video containing 98 distinct ageist remarks.

5.3.2 Comparative analysis of ageist and non-ageist comments

To contextualize ageist discourse, we compared ageist comments (N = 810) with non-ageist comments (N = 3579) across toxicity, profanity, insult, engagement, and sentiment scores. The comparison reveals statistically significant differences between the two groups on several measures.

Toxicity, Profanity, and Insult: Ageist comments are consistently more toxic, profane, and insulting than non-ageist comments. The mean toxicity score for ageist comments is 0.168, higher than the 0.135 mean for non-ageist comments (Mann-Whitney U $p < 10^{-6}$, t-test $p = 0.0036$). Similarly, profanity and insult scores are elevated in ageist comments (profanity means: 0.024 vs. 0.014, insult means: 0.081 vs. 0.065), with highly significant differences (profanity Mann-Whitney U $p < 10^{-5}$, insult Mann-Whitney U $p = 0.00034$). These findings indicate that ageist remarks on YouTube tend to be more harmful and offensive, even though median values for both groups are close to zero, suggesting that only a subset of comments exhibits strongly toxic or profane language.

Engagement: Non-ageist comments receive more engagement than ageist ones. The average like count for non-ageist comments is 8.86, compared to 3.39 for ageist comments, although both groups share a median of zero likes (t-test $p \approx 1.37 \times 10^{-7}$, Mann-Whitney U $p = 0.585$). This suggests that, while some ageist comments attract attention, non-ageist content is more likely to generate broader positive engagement within the community. No significant differences are observed in the number of replies per comment.

Sentiment: Ageist comments are more negative than non-ageist comments according to VADER sentiment analysis (mean: -0.183 for ageist, -0.022 for non-ageist; Mann-Whitney U $p < 10^{-11}$, t-test $p < 10^{-13}$). However, TextBlob scores indicate only a slight difference in mean sentiment (0.066 for ageist vs. 0.055 for non-ageist; differences not statistically significant). This highlights that ageist discourse is more likely to be accompanied by negative emotion and tone, reinforcing stereotypes and negative attitudes towards older adults.

5.3.3 Longitudinal (event-centered) analysis

To understand how key pandemic milestones influenced the prevalence of ageist content, we conducted an event-centered longitudinal analysis of the proportion of ageist comments over time. We selected four major COVID-19-related events and examined the mean proportion of ageist comments before and after each event (see Fig. 8).

- *FDA Approves Emergency Use for the Vaccine*[8]: The mean proportion of ageist comments increased from 0.1702 before the announcement to 0.2026 after (p-value = 0.0877), suggesting a potential uptick in ageist discourse following this major vaccine milestone, although the difference is not statistically significant.

[8] https://www.cdc.gov/museum/timeline/covid19.html.

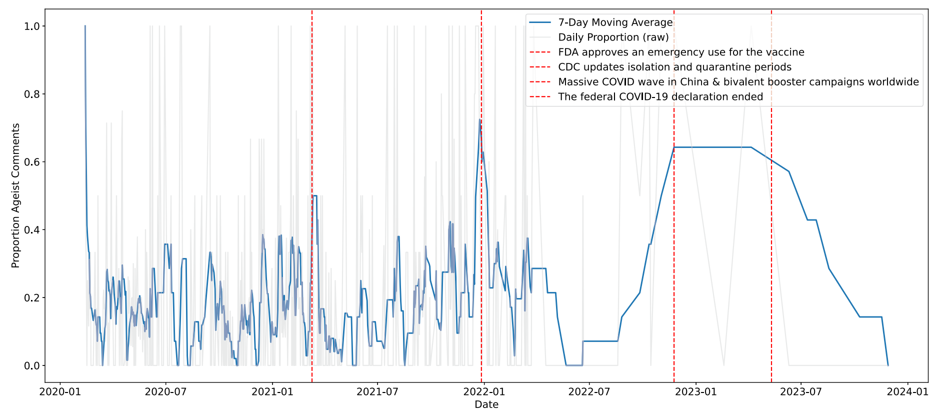


Figure 8 Proportion of ageist comments over time, with major COVID-19-related events indicated

- *CDC Updates Isolation and Quarantine Periods* [9]: The proportion rose from 0.1784 pre-event to 0.2245 post-event (p-value = 0.3793), again with no statistically significant change but with a numerical increase.
- *Massive COVID Wave in China and Bivalent Booster Campaigns*[10]: Ageist comment rates moved from 0.1843 to 0.2000 (p-value = 0.2745), indicating a modest but non-significant shift.
- *End of Federal COVID-19 Declaration*[11]: Here, we observe a significant drop, with the mean proportion decreasing from 0.1869 pre-event to zero post-event (p-value = 0.0208). This suggests that the official conclusion of the public health emergency may have reduced ageist discussion in online discourse.

As illustrated in Fig. 8, the temporal trend is visualized using both the daily raw proportion of ageist comments and a 7-day moving average. While fluctuations and short-term increases follow certain major events, only the conclusion of the federal emergency is associated with a statistically significant reduction in ageist commentary.

This analysis highlights the value of event-centered longitudinal approaches for identifying how ageist sentiment may respond to public health policy shifts, pandemic milestones, and global events. It also demonstrates that while some events may generate temporary increases in ageist discourse, the formal end of the emergency is linked with a notable decline.

6 Discussion

In this paper, we introduce AGECovP, a novel dataset designed for examining discussions about older adults on YouTube during the initial phase of the COVID-19 pandemic, covering a range of topics. To the best of our knowledge, this dataset represents the first of its kind for investigating the portrayal, perceptions, and conversations related to older adults through social media data. Recognizing that older adults are among the populations at increased risk, particularly during pandemics, AGECovP provides valuable insights into the societal and digital narratives that shape their well-being. By leveraging this dataset,

[9] https://news.nus.edu.sg/looking-to-2022-what-lies-ahead-in-the-covid-19-pandemic/.

[10] https://news.nus.edu.sg/looking-to-2022-what-lies-ahead-in-the-covid-19-pandemic/.

[11] https://archive.cdc.gov/www_cdc_gov/coronavirus/2019-ncov/your-health/end-of-phe.html.

researchers can better understand the vulnerabilities and systemic biases older adults face in digital spaces and how these factors may contribute to inequalities.

AGECovP offers extensive opportunities to explore how old age is perceived on social media and to analyze the representation of older individuals through video content. While the video content predominantly consists of media materials that deliver supportive and informative messages, the accompanying comments reveal a contrasting perspective. These comments often highlight negative and toxic discourse, including conspiracy theories, misinformation, age-based prejudice, and discrimination. Additionally, we provide labeled ageist content, which serves as a valuable resource for understanding and detecting age-based bias in online discussions. AGECovP highlights the importance of understanding the multifaceted nature of online discussions and how they can influence the well-being and societal perceptions of older adults, especially in times of crisis.

In this regards, first, we hope that the dataset can be used for further exploring societal perceptions of aging by examining the content and interactions related to older adults on the YouTube platform. Specifically, the most dominant topic ("Elderly Health and Societal Challenges")'s association with negative sentiment and expressions of isolation provides a foundation for exploring the psychological impact of pandemic-related measures on older adults. Research can investigate the emotional toll of new regulations and the implications for mental well-being.

Second, the prevalence of claims suggesting intentional creation of COVID-19 to eliminate the older adults and assigning blame to older adults for pandemic regulations invites further investigation. Future studies can focus on the spread of misinformation, the origins of such narratives, and their impact on public perceptions. This is particularly promising since one of the most prevalent topics in the AGECovP comments was about misinformation and conspiracy theories.

Third, we hope the dataset helps researchers understand the anti-vaccine movement, particularly as it pertains to older adults, as highlighted by the results from the social network analysis. For instance, researchers can delve into the users of the "anti-vaccine movement" cluster to identify key influencers, recurring themes, and patterns within anti-vaccine discussions about older adults. Further, given the dataset's focus on older adults, researchers can specifically examine how older adults are framed within the anti-vaccine discussions. For example, are there specific factors utilized by the anti-vaccine movement that may influence vaccine hesitancy among older adults, as reflected in the discussions within the dataset? The findings from such analysis can have implications for public health policy.

Forth, we anticipate that the dataset will serve as a valuable resource for examining ageist content in online discussions, providing insights into the ways older adults are depicted and perceived. The labeled ageist comments within AGECovP provide a foundation for examining the prevalence, context, and patterns of age-based discrimination and prejudice. Researchers can analyze these comments to identify common stereotypes, recurring themes, and influential users perpetuating ageist narratives. Additionally, they can explore the factors that contribute to ageist discourse and assess the potential impact on public attitudes toward older adults. By investigating how ageism manifests in social media discussions, researchers can uncover underlying social biases and inform strategies to combat age-based discrimination. Insights from such studies can guide interventions aimed at promoting respect and inclusivity for older adults, ultimately influencing public awareness campaigns and policy development.

Beyond pandemic-related research, AGECovP's rich metadata and annotated comments enable researchers to address a range of concrete questions on aging and online discourse. For example, by leveraging the TextBlob and VADER sentiment scores, researchers can compare how older adults are discussed in videos from different countries, shedding light on cultural differences in the framing of aging. The Category and Tags fields allow for the analysis of themes such as elder care or digital literacy—enabling studies of how frequently topics like retirement or assisted living appear and whether associated comments contain ageist language. By joining comment-level data with video engagement metrics, it is possible to explore whether videos that contain higher levels of ageist content also receive more user interaction, or if toxic age-based narratives are amplified within highly viewed channels. The dates and duration information can be used to examine whether ageist discourse intensifies around major events (such as policy changes or health crises) or whether longer videos foster more nuanced conversations. In addition, the channels' information enables analysis of repeated user behavior, such as whether certain users are more likely to post ageist comments, which can inform strategies for community moderation or digital literacy interventions. Taken together, this information make AGECovP a versatile resource for specific, data-driven investigations into ageism, online health narratives, policy discourse, and the social dynamics of digital aging.

One limitation of the AGECovP dataset comes from its inclusion of only 1st level comments, a constraint imposed by YouTube API limitations. Future efforts aim to overcome this by exploring alternative datasets for a more comprehensive analysis. Additionally, the reliance on keyword searches to identify content related to older adults and COVID-19 presents another limitation. To address this, future plans involve implementing a classifier for more efficient and targeted identification of relevant content.

Another limitation of the AGECovP dataset lies in its scope, as it primarily captures the perspectives of users who are willing to express their opinions online. This excludes opinions shared offline or by individuals who prefer not to use YouTube as a platform for idea dissemination. Additionally, our study is constrained by the moderation policies enforced by the YouTube platform. Since we focus solely on content permitted within these guidelines, any material removed or deleted due to violations of platform rules cannot be studied. As a result, our analysis may not encompass the entire content related to COVID-19 and older adults, as it remains contingent on the platform's regulations and the content made available for analysis.

In future work, we plan to conduct a qualitative analysis of the labeled ageist content to better understand the narratives and themes surrounding ageism in social media discussions. By examining the context and underlying discourse, we aim to uncover the subtle ways in which ageist attitudes are conveyed and reinforced. This approach will provide deeper insights into the types of language and rhetoric used to marginalize older adults and help illuminate the broader implications of such content in shaping public perceptions and attitudes. Understanding these narratives can guide the development of targeted interventions and policies aimed at reducing ageist sentiments online.

Further, it is important to note that while tools like VADER and TextBlob offer reliable sentiment classification for most direct language, they have limited capacity to detect sarcasm or comments where surface-level positive language conveys negative intent. Manual review of a sample of misclassified comments indicated that sarcastic content was often assigned incorrect sentiment scores. Addressing sarcasm and similar phenomena remains a challenge for automated sentiment analysis and is a potential avenue for future methodological improvements in analyzing social media discourse.

Finally, all engagement metrics (such as views, likes, and comments) represent a single snapshot from the time of data collection and are not updated in real-time. In future work, the dataset could be refreshed to reflect the latest engagement statistics and provide insight into how interactions evolve over time.

6.1 FAIR consideration

The provided dataset adheres to the FAIR principles of Findability, Accessibility, Interoperability, and Reusability. Access to the dataset is available on Zenodo via the DOI: https://zenodo.org/records/15800324, providing a stable and citable repository. AGECovP is released as a static resource, representing a comprehensive snapshot of YouTube videos and comments related to older adults during the initial phase of the COVID-19 pandemic. The dataset is not intended to be continuously updated, and no ongoing updates or additions are planned. By depositing the data on Zenodo, we ensure long-term preservation and open access, allowing researchers worldwide to reuse and analyze the dataset in accordance with Zenodo's terms of service. The dataset is not affiliated with or maintained by a specific university or institution; rather, it is managed as a community resource for research on aging, social media, and online discourse.

6.2 Ethical consideration

We carefully crafted our dataset during the data collection phase, focusing exclusively on publicly accessible information on YouTube through the YouTube's Data API. In particular, to protect the anonymity of channels, actual author ids are anonymised, but they can be recollected using the YouTube Data API for future research purposes. Additionally, our dataset collection methodology has been granted exemption from ethics review by Toronto Metropolitan University in North America's Research Ethics Board (REB).[12] This exemption is based on the following criteria: (1) research activities that do not entail direct interaction with individuals and where individuals do not anticipate privacy concerns; (2) utilization of datasets for analysis that are either publicly available or protected by law.

It is important to note that, similar to other datasets in this domain, AGECovP dataset carries the potential for misuse and negative societal impacts. Specifically, the dataset contains content linked to misinformation, stereotypes, and conspiracy theories. In case of misuse, improper analysis, or biased interpretation, there is a risk of spreading misinformation, reinforcing stereotypes, and supporting unfounded conspiracy theories related to COVID-19, potentially contributing to public confusion and fear. To mitigate these risks, we provide transparent methodologies used for the data collection and analysis. We further discuss the dataset's limitations and appropriate contextualization of findings with the hope to minimize negative societal impacts.

Abbreviations

AGECovP, Aging Perceptions in COVID-19 Pandemic; API, Application Programming Interface; CCP, Chinese Communist Party; COVID-19, Coronavirus Disease 2019; DNC, Democratic National Committee; RTs, Retweets; SBERT, Sentence-BERT;

[12] https://www.torontomu.ca/research/resources/ethics/course-based-research-ethics/#!accordion-1636735697402-what-does-not-require-review.

UCC, User Copied Content; UGC, User Generated Content; URL, Uniform Resource Locator; VADER, Valence Aware Dictionary and sEntiment Reasoner.

Acknowledgements
Not applicable.

Author contributions
A.G designed the study and wrote the manuscript. N.K and S.A performed the analyses.

Funding information
Partial financial support was received from Mitacs Globalink Research Internship (GRI) as well as the Natural Sciences and Engineering Research Council of Canada (NSERC) (RGPIN-2022-03809).

Data availability
Full access to the original data available at the repository: https://zenodo.org/records/14247253.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Ted Rogers School of Management, Toronto Metropolitan University, Toronto, Canada. [2]School of Computing, National University of Singapore, Kent Ridge, Singapore. [3]Department of Computer Science, University of Western Ontario, London, Canada.

Received: 2 December 2024 Accepted: 19 August 2025 Published online: 27 August 2025

References

1. Bambra C, Riordan R, Ford J, Matthews F (2020) The covid-19 pandemic and health inequalities. J Epidemiol Community Health 74(11):964–968
2. Armitage R, Nellums LB (2020) Covid-19 and the consequences of isolating the elderly. Lancet Public Health 5(5):256
3. (2021) United Nations: pandemic exposes ageism and age discrimination in society, says UN expert. https://www.ohchr.org/en/press-releases/2021/09/pandemic-exposes-ageism-and-age-discrimination-society-says-un-expert
4. Wister A, Speechley M (2020) Covid-19: pandemic risk, resilience and possibilities for aging research. Can J Aging 39(3):344–347
5. Barrett AE, Michael C, Padavic I (2021) Calculated ageism: generational sacrifice as a response to the covid-19 pandemic. J Gerontol, Ser B, Psychol Sci Soc Sci 76(4):201–205
6. Lagacé M, O'Sullivan T, Dangoisse P, Mac A, Oostander S, Doucet A (2022) A case study on ageism during the COVID-19 pandemic. Employment and Social Development Canada, Ottawa
7. Fraser S, Lagacé M, Bongué B, Ndeye N, Guyot J, Bechard L, Garcia L, Taler V, Adam S, et al (2020) Ageism and covid-19: what does our society's response say about us? Age Ageing 49(5):692–695
8. Gruzd A, Mai P (2022) Communalytic: a research tool for studying online communities and online discourse
9. Albadi N, Kurdi M, Mishra S (2019) Hateful people or hateful bots? Detection and characterization of bots spreading religious hatred in Arabic social media. In: Proceedings of the ACM on human-computer interaction 3(CSCW), pp 1–25
10. Vrijenhoek S, Bénédict G, Gutierrez Granada M, Odijk D, De Rijke M (2022) Radio–rank-aware divergence metrics to measure normative diversity in news recommendations. In: Proceedings of the 16th ACM conference on recommender systems, pp 208–219
11. Cambria E, Li Y, Xing FZ, Poria S, Kwok K (2020) Senticnet 6: ensemble application of symbolic and subsymbolic ai for sentiment analysis. In: Proceedings of the 29th ACM international conference on information & knowledge management, pp 105–114
12. Cheng J, Bernstein M, Danescu-Niculescu-Mizil C, Leskovec J (2017) Anyone can become a troll: causes of trolling behavior in online discussions. In: Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing, pp 1217–1230
13. Ding Y, Du Y, Hu Y, Liu Z, Wang L, Ross K, Ghose A (2011) Broadcast yourself: understanding youtube uploaders. In: Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference, pp 361–370
14. Reimers N, Gurevych I (2019) Sentence-bert: sentence embeddings using Siamese bert-networks. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, pp 3982–3992. https://aclanthology.org/D19-1410/
15. Loria S, et al (2018) Textblob documentation. Release 0.15 2(8):269
16. Hutto C, Gilbert E (2014) Vader: a parsimonious rule-based model for sentiment analysis of social media text. In: Proceedings of the international AAAI conference on web and social media, vol 8, pp 216–225
17. Viera AJ, Garrett JM, et al (2005) Understanding interobserver agreement: the kappa statistic. Fam Med 37(5):360–363

18. Bonta V, Kumaresh N, Janardhan N (2019) A comprehensive study on lexicon based approaches for sentiment analysis. Asian J Comput Sci Technol 8(S2):1–6
19. Hanu L (2020) Unitary team: detoxify. Github. https://github.com/unitaryai/detoxify
20. Xu W (2021) Ageism in the media: online representations of older people. PhD thesis, Linköping University Electronic Press
21. Ng R, Indran N, Liu L (2022) Ageism on Twitter during the covid-19 pandemic. J Soc Issues 78(4):842–859
22. Levy BR, Chung PH, Bedford T, Navrazhina K (2014) Facebook as a site for negative age stereotypes. Gerontologist 54(2):172–176
23. Iversen TN, Larsen L, Solem PE (2009) A conceptual analysis of ageism. Nord Psychol 61(3):4–22
24. Landis J (1977) The measurement of observer agreement for categorical data. Biometrics
25. Fleiss JL (1971) Measuring nominal scale agreement among many raters. Psychol Bull 76(5):378

Publisher's note

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

Submit your manuscript to a SpringerOpen journal and benefit from:

- ► Convenient online submission
- ► Rigorous peer review
- ► Open access: articles freely available online
- ► High visibility within the field
- ► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com