methods, data, analyses | 2025, pp. 1-32 DOI: 10.12758/mda.2024.08

Augmenting Data Download Packages – Integrating Data Donations, Video Metadata, and the Multimodal Nature of Audio-visual Content

Lion Wedel[1], Jakob Ohme[1] & Theo Araujo[2]

[1] *Weizenbaum Institute for the Networked Society, Germany*

[2] *Amsterdam School of Communication Research, Netherlands*

Abstract

This research explores the potential of augmented Data Download Packages (aDDPs) as a novel approach to analyze digital trace data, using TikTok as a use case to demonstrate the broader applicability of the method. The study demonstrates how these data packages can be used in social science research to understand better user behavior, content consumption patterns, and the relationship between self-reported preferences and actual digital behavior.

We introduce the concept of aDDPs, which extend the conventional Data Download Packages (DDPs) by augmenting the collected data with survey data, metadata, content data, and multimodal content embeddings, among other possibilities - rendering aDDPs an unprecedentedly rich data source for social science research. This work provides an overview and guidance on collecting, augmenting DDPs, and analyzing the resulting aDDPs.

In a pilot study on 18 aDDPs, we use the combination of data components in aDDPs to facilitate research on user engagement behavior and content classification. We showcase the potential of the information breadth and depth that aDDPs depict by exploiting the combination of multimodal content embeddings, the users' watch history, and survey data. To do so, we train and compare uni- and multimodal classifiers, classify the 18 aDDPs' videos, and investigate the extent to which user engagement behavior impacts future content suggestions. Furthermore, we compare the users retrieved content with the users' self-reported content consumption.

Keywords: data download packages, augmentation, multimodality, TikTok, vertical videos, classification

TikTok is one of the fastest-growing social media platforms worldwide (Newman et al., 2023). In addition, its role in distributing information during the COVID-19 crisis and the Russian invasion of Ukraine, as well as the discussions around its Chinese ownership, manifests the understanding that the platform needs to be considered relevant for social media researchers of many fields (e.g., Basch et al., 2020; Primig et al., 2023). The European Commission has recently recognized this relevance, assigning TikTok the status of a very large online platform (VLOP), which can carry systemic risk for the European Union (*DSA*, 2023). As a vertical video platform (VVP), TikTok's main characteristics are short vertical videos (recorded in portrait mode) and the substantial reliance on algorithmic curation and passive use compared to other social media platforms (Hase et al., 2022). Unlike Twitter or Facebook, TikTok content is inherently multimodal beyond text and an occasional picture – consisting of audio-visual information. This creates new challenges and opportunities for computational social sciences and adjacent fields.

The EU General Data Protection Regulation (GDPR, 2016) allows users to demand the data TikTok has collected about them (TikTok, 2023b). Similar laws exist in countries and regions beyond the EU, such as Japan or Brazil (Boeschoten et al., 2020). The access explicitly allows sharing data with *"…third parties, such as social scientists."* (ibid., p. 4). This is the foundation to explore the potential of data donations for user-centered research purposes. Still, research utilizing Data Download Packages (DDPs) from video platforms like TikTok is sparse, given the expected difficulties of retrieving and analyzing the multimodal nature (i.e., moving images, audio, and text) of (vertical) videos. Specifically, it is difficult for social science research to understand exposure patterns based on data donations. It is, therefore, essential to develop new approaches to understand the content that, within the EU alone, around 135.9 million users are exposed to monthly (TikTok, 2023a).

This paper explores the potential of augmented DDPs (aDPPs) for social science researchers to study information exposure and conduct algorithmic auditing on TikTok. It presents a new approach, integrating TikTok DDPs with 1) survey data, 2) video metadata, 3) content data, and 4) the multimodal features of a TikTok post. Previous research has identified multiple challenges to arrive at a meaningful basis for social science research that allows the analysis of vertical video platform exposure data with DDPs (Boeschoten et al., 2021; Driel et al., 2022; Ohme et al., 2021). While we leave some of those unaddressed (e.g., sample biases and conversion rates of successful donation), we describe two challenges on the way to an augmented TikTok data download package: 1) the data donation

Direct correspondence to
Lion Wedel, Weizenbaum Institute for the Networked Society, Berlin, Germany
E-mail: lion.wedel@weizenbaum-institut.de

process and 2) the augmentation of DDPs. Subsequently, we provide solutions for tackling the described challenges in a pilot study. The concept of aDDPs is not limited to TikTok. It can serve as a guiding concept for research using data donations from any social media and content platform where the native DDP does not hold sufficient information to answer the proposed research questions.

In the following, we will first explain the background and relevance of the topic before we explain how TikTok data donations can be augmented with specifically multimodal content features. In the last exploratory part, we show how aDDPs can be used in social science research to answer substantial questions, such as how previous engagement affects future suggested content and whether user perceptions of their information consumption align with the empirical findings.

TikTok's Inherent Multimodality and the Potential of aDDPs

Over the last decade, multimedia content has increased in importance in delivering media messages to users and audiences. In this context, muti-modality describes the combination of different modes of content, such as *"... language, images, typography [or] layout ..."* in a media format (Hiippala, 2017, p.421). Since their emergence, text and still images have been the predominant modes of content presentation on digital platforms, often in separated elements. With vertical video features such as Instagram Stories, Snapchat Spotlight, YouTube Shorts, and TikTok as the dominant vertical video-only platform, moving image is combined with audio tracks. This multimodality is further enhanced by integrating still images, icons, and text, such as hashtags or subtitles. This integration of different content modes in the format of a video challenges existing media analysis paradigms (e.g. Valkenburg, 2022) and calls for new approaches to preparing multimodal content for analysis. TikTok's platform logic is based on videos with audio and a description – thereby inherently multimodal (Hase et al., 2022).

Social scientists have a clear interest in researching video-only platforms such as TikTok but often retreat methodologically to qualitative methods (e.g., Mordecai, 2023; Zhou Ting, 2021), especially considering the complexity of multimodal data. Here, a set of contributors usually manually labels a sample of videos (e.g., Li & Kang, 2023; Ming et al., 2023; Ng & Indran, 2023; Yeung et al., 2022). Labeling posts for social science research aligns with a classification task in machine learning. Hence, the collection of DPPs and their augmentation are the first two steps. In the third step, a large-scale classification model is necessary to unfold the potential of aDDPs for critical research social scientists seek concerning video-only platforms.

A handful of contributions acknowledge the multimodality of TikTok videos, and the consequent contribution to the development of uni- and multimodal classifiers must be mentioned here. With *SexTok,* George & Surdeanu (2023) present a 1,000 video dataset on which they train separately a text and a video embedding-based classifier to predict one of three classes. Other pieces on TikTok videos extract text shown within the video or focus only on the audio feature to classify videos subsequently (e.g., Fiallos et al., 2021; Ibañez et al., 2021). Such work relies on one modality, ignoring the information depth other modalities could add. Kim et al. (2023) embrace TikTok as a multimodal platform but eventually reduce videos to thumbnails and audios to transcripts – in both cases, scrutinizing the information depth those modes might entail. Nevertheless, they showcase that using variables retrieved through pre-trained classifiers as the basis for scalable classification and subsequent analysis, such as hypothesis testing, is a feasible approach for research on TikTok and possibly other video-only platforms.

Research across domains has consistently shown that incorporating all available modalities improves the performance of classification tasks (e.g., Pandeya & Lee, 2021; Qi et al., 2023; Shang et al., 2021). Specifically in the application of social media posts, multimodal approaches have been proven to equalize weaknesses of unimodal representations in Instagram posts (Zeppelzauer & Schopfhauser, 2016). A truly multimodal classification approach to TikTok videos is presented by Shang et al. (2021), who take visual content, audio, video descriptions, and engagement data into account. It is trained and tested on 226 misleading and 665 non-misleading videos. However, they do not report on the performance of unimodal or non-neural network approaches – not ruling out that a multimodal neural network approach might be unnecessary. A comparison of different methods and modalities for the classification of fake news on TikTok is provided with *FakeSV* (Qi et al., 2023). They offer significant first evidence for the usefulness of multi-modal classification of Chinese (fake)-news TikTok videos.

A caveat for previous research is that they are trained and tested on datasets collected via hashtag, author, or event lists and/or are being hand-curated from the beginning. Those datasets only reflect a subset of the variety of videos users are possibly exposed to on TikTok. The classifiers trained on such data might not allow for a reliable classification of datasets that contain increased content variability, such as actual user trace data.

Collecting videos via a hashtag, keyword, or actor sample might tell us something about those topics and actors (and can serve to train a classifier). Still, it hardly tells us anything about the exposure to or impact of such content - what users consume and to what extent. Here, *data donations* present an excellent approach to gathering user-centric data that gives researchers access to watched videos. Two recent studies on TikTok base their findings on TikTok DDPs. They dive into analysis based on the raw DDPs and an accompanying survey, leav-

ing questions of content exposure and multimodality unexplored (Goetzen et al., 2023; Zannettou et al., 2023). Hence, research has yet to use the full potential of TikTok DDPs to analyze exposure to multimodal content. However, the lack of understanding of content exposure and the multimodal nature of TikTok have posed two challenges for research on TikTok: 1) the facilitation of collecting TikTok DDPs and 2) the augmentation of said DDPs.

While we focus on the case of TikTok in this paper, the description also holds for digital platforms that are similarly multi-modal and have a vertical video feature, such as YouTube (Shorts) and Instagram (Reels). For those, an augmentation step is necessary for research incorporating the content level since the DDPs only contain metadata (Driel et al., 2022). For text-heavy platforms such as Facebook or Twitter, the DDPs already contain bigger parts of the content. However, these DDPs can, for example, be augmented with the full texts of articles users click on or post about. aDDPs are, hence, a generalizable approach that aims to increase the depth of available data for analysis, combining data not included in DDPs and corresponding survey data.

Challenge 1: The Data Donation Process & Available Frameworks

Digital trace data can be roughly differentiated into platform-centric and user-centric data. Platform-centric data is mainly gathered via APIs (often, this is publicly available data collected retrospectively without explicit user consent), while user-centric data is gathered either through tracking approaches on user devices (prospectively) or via data donations (Ohme et al., 2023). For TikTok, APIs or web scraping do not provide user-centric data. While they provide public data, private information such as the user's watch history and their behavior around each video is beyond their capabilities. Here, DDPs are the best option for collecting user-centric data to explore content exposure and the behavior of users.

DDPs provide an ecologically valid, non-reactive, reasonably scalable, and geographically independent data source – a combination of traits that no other user-centric data collection method provides (Driel et al., 2022; Ohme et al., 2023). DDPs represent the most complete available collection of user-centered digital trace data from TikTok available to date. Importantly, DDPs from TikTok give, at the time of our data collection in August 2023, the link to each video that was watched by a user - allowing for retrospective[1] data augmentation and making TikTok DDPs especially valuable for digital communication research (ibid.).

1 Our current data collection has shown that the watch history contained in the DDPs only dates back half a year from the point of the data request. Other activities such as liking, commenting and private messages are present for the whole time of an account's existence.

To collect TikTok DDPs, a user must request the data as a JSON or TXT file and donate their data. The resulting *Data Download Package* (DDP) is a set of user-centric digital trace data (Ohme et al., 2023). The data donation, in general, can be facilitated in three different ways: First, researchers instruct the participants to install a desktop or mobile application that performs preprocessing steps locally and then sends the final DDP to the researchers' server (e.g, DataSkop, 2023). Second, researchers instruct the participants to upload the data directly to a server under their control, only to conduct data privatization and minimization afterward (e.g., Driel et al., 2022) or, third, use a web-based application that executes preprocessing steps on the participant's local machine, thereby only saving the final DDP to the researchers' database (e.g., Araujo et al., 2022; Boeschoten et al., 2023; Friemel & Pfiffner, 2023).

For the collection of TikTok DDPs, the third approach is ideal. It has the advantage of running the preprocessing locally, and current web applications are platform-independent – making the donation as easy and safe as possible for participants. Compared to the other two approaches, the threat of *compliance & consent error* (see Boeschoten et al., 2020) is mitigated as much as possible - *compliance* in the case of a dedicated desktop app that has to be installed and needs the user to transfer data between devices and *consent* in case of the direct data transfer – demanding the participant to donate not just the data required by researchers but also data such as address, name and personal messages. With *Port* (Boeschoten et al., 2023) and *DDM* (Pfiffner et al., 2022), at least two frameworks for a web app with the described advantages are in development and partially already published under open-access licenses to be used by researchers – the future of data donations is thereby set on web applications that allow for maximum privacy by minimal inconvenience for the donor.

For the current study, *Port* was employed, which allows for preprocessing on the participant's device, thereby mitigating privacy concerns for participants. Participants were recruited through a convenience sample, with a call for participation distributed via colleagues and student courses. Participants were initially led to an online survey that collected sociodemographic data and contained questions about their perception of the content they received on TikTok (further described in the section "Applying aDDPs in TikTok"). The survey also included detailed instructions on how to request their DDP from TikTok. During the survey, we generated a unique ID for each participant to link the survey data and the data donation. During the study, TikTok took up to three days to prepare the file (TikTok, 2023b). After three days, participants received an E-Mail with a personalized (via the ID) link to *Port*, where they found a manual on uploading their data donations. The ID is saved along with the data donations, allowing us to connect the survey data and the data donations later. 18 out of 42 (42.68%) recruited participants completed the process. Participants received an incentive of 20 € upon completion. The study received approval from the Ethical Review

Boards of the Weizenbaum Institute and the University of Amsterdam. An overview of the included information in the locally processed and donated DDPs can be found in Table 1.

While the data package that researchers retrieve is often only a subset of the DDP that the user has downloaded (depending on the preprocessing), we will continue to describe the donated data package as a (augmented) data download package since the subset that is augmented represents one to one the user trace data of the respective activities contained in the DDP (e.g., watch history).

Table 1 Description and collected variables for each activity beyond the timestamp. The timestamps always mark the beginning of the respective activity.

Activity	Description	Additional variables collected
Following	The user is following another user.	-
Favorites	The user is marking a video as a favorite.	Link to video
Logging in	The user is logging into their TikTok account.	Operating System
Searching	The user is searching TikTok with a search term.	-
Sharing	The user is sharing the present video in-app or externally.	-
Watching Videos	The user is watching a video.	Link to Video
Blocking	The user is adding another user to the block list.	-
Commenting	The user is commenting on a video.	-
Chatting	The user is writing a private message to another user.	-
Going Live	The user is starting a live stream.	-
Watching Livestreams	The user is watching a live stream	Link to Video
Posting Videos	The user is posting a video of their own.	Likes
Liking	The user is liking a video.	-

Challenge 2: Augmenting TikTok DDPs

TikTok DDPs provide a variety of insightful data points such as user activities (liking, sharing, watching), the users' app settings, and ad interests (Zannettou et al., 2023). Research has described different ways of linking DDPs with other data sources like survey data (e.g., Haim et al., 2023; Stier et al., 2020) and scraped metadata (e.g., video length, likes - Goetzen et al., 2023; Zannettou et al., 2023). Our suggested approach goes one step further. It proposes integrating audio-visual content features and their machine-readable multimodal feature embeddings (also multimodal representations) to the TikTok DDP, video metadata & survey data (see Figure 1). A resulting augmented data download package (aDDP) contains *survey data*, the *donated (subset) of* the *data download package*, *metadata* of a post (such as video length or number of likes), *content data* of a post (such as the video and audio file), and finally, *multimodal representations* of each post. These, ultimately, can serve as input for subsequent supervised and unsupervised machine learning tasks. Such aDDP combines the advantages of collecting initial user-centric data via DDPs with the richness of publicly accessible metadata and analyzable audio-visual content features. We do not stop with the augmentation via established computational methods in the social sciences (metadata scraping, natural language processing) but exploit the full depth of audio-visual content to facilitate state-of-the-art research. The concept of aDDPs provides a terminology that covers data linkage efforts and combines them with the advanced methodological opportunities of contemporary computational research.

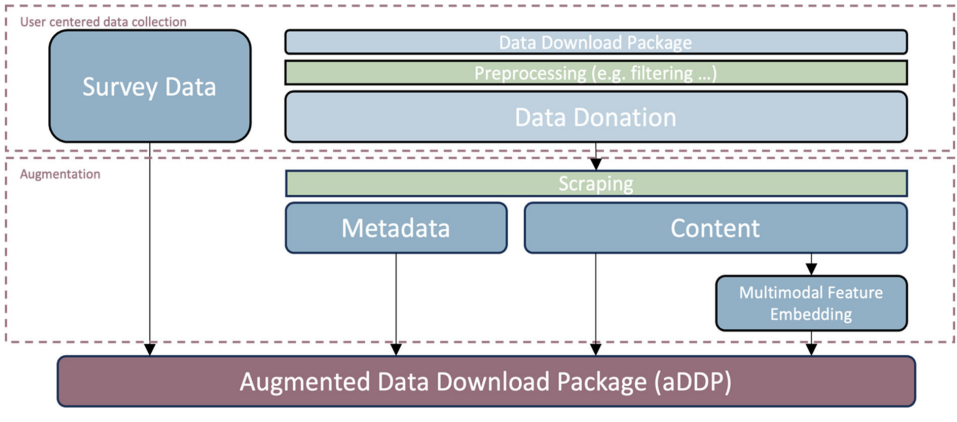


Figure 1 Process of Data Donation Augmentation.

The augmentation with a survey during the initial data collection (e.g., Haim et al., 2023; Stier et al., 2020), as well as the initial collection of TikTok DDPs (e.g., Goetzen et al., 2023; Zannettou et al., 2023) itself, has been discussed previously; the other steps of the augmenting process demand a more detailed description and reflection to guide the present, and future research. Hence, in this paper, we focus on collecting metadata and content data and specifically explain the multimodal feature extraction for TikTok. This process, however, can be helpful in different projects in social science research that deal with multimodal content. To provide such guidance in an appropriate form, we will now go through the methodological decisions of the augmentation process. The substeps are exemplified with a pilot study of 18 data donors, showcasing the possibilities for empirical research based on aDDPs.

Collecting Meta and Content Data for TikTok

The TikTok Research API is not viable for our purpose because it only provides minimal metadata and no video or audio files – making other data sources necessary (Meßmer et al., 2023). At the same time, the terms of services forbid any other way of data augmentation in the case of using the API (TikTok, 2023). We thereby choose not to use the TikTok API.

Alternative public Python packages can facilitate the scraping instead, returning many more variables than the TikTok Research API, such as the *Pyktok* or *TikTok-Api* packages (Freelon, 2022/2023; Teather, 2019/2023). We found using the *TikTok-Api* package to be sufficiently reliable and convenient for data collection. To download the videos, we used a custom Python script. With the videos downloaded, the audio can be extracted with, e.g., the open-source Python package *moviepy* (Zulko, 2013/2023). For further information on the usage of the mentioned packages, please refer to their documentation.

In sum, the augmentation step of retrieving metadata and video data can currently not be sufficiently facilitated without programming and web scraping knowledge. As it comes with unofficial and custom scrapers, the scraping is volatile due to changes in website architecture. Custom scraping also poses a challenge to time management – because of its slowness and unreliability. Finally, scraping of content from the web poses legal questions. However, we deem our research in line with current EU legislation.[2]

An unsolvable circumstance of the current affordances for metadata and content data scraping is that we can not retrieve data for posts that are no longer available – be it for violations against the platforms' terms of service or the users' changed privacy settings. In our case, at the time of scraping, we could no

2 The research is carried out by a non-profit research institute with the primary goal of scientific research. The scraping thereby falls under the exception granted by the DSM Directive for text and data mining. (Egger et al., 2022 p. 73-75)

longer retrieve data for 13.58% of the videos from the analyzed sample (1,821 out of 13,342 videos), which is similar to previous research on TikTok data donations (Zannettou et al., 2023).

Extract Feature Embeddings from TikTok Data

Depending on the research question and domain, the audio-visual content itself (e.g., manual content analysis) and meta-data can be used directly for analysis as they are. For machine learning tasks, there are two main options for representing the modalities: 1) using technical content characteristics such as cutting frequency or color spectrum (visual) and the loudness or dynamic complexity (audio) (e.g., Huddar et al., 2020; Ibañez et al., 2021; Lepa & Suphan, 2019; Syed et al., 2021) or 2) a vector representation retrieved from a pre-trained general-purpose model of the gathered modalities (e.g., Chiatti et al., 2019; Ram et al., 2020; Reeves et al., 2021). The latter approach (*transfer learning*) is at least equally good, often better for follow-up classification tasks compared to embeddings based on technical characteristics (Baltrusaitis et al., 2019; Zhang & Peng, 2022). This can be explained by them not being bound by the researchers' assumptions and knowledge of the possibilities surrounding each mode (Qi et al., 2023). Instead, the complexity of their training data binds the pre-trained models used to retrieve the embeddings. A typical training dataset for video representations is the *Kinetics 400* – a dataset that returns a vector of length 400 reflecting 400 human actions within the videos (Kay et al., 2017). The final layer and, even more, the last hidden layer – commonly larger and less impacted by the model's training classes - can be assumed to hold a sufficient number of latent characteristics of a video (or any other input modality) – superseding any hard-coded assumption made by the researchers.

The choice of how to retrieve the feature embedding is a core aspect of a multimodal classification task (Sleeman et al., 2021). Unlike in computer sciences, the models used to generate the embeddings should not merely be assessed based on their performance (Bender et al., 2021; Schwartz et al., 2020). When applied in the context of computational social science, the ease of implementation of a model becomes a significant factor. Since the performance difference between easily accessible pre-trained models and newer models that might be too recent to be accessible is usually in the lower one-digit percentages. Thus, the performance gain does not justify the added time spent on the implementation. Therefore, we suggest utilizing models that are easily importable in major machine learning libraries like *PyTorch* or *TensorFlow*. Both facilitate a hub of pre-trained models (*TensorFlow Hub*[3], *PyTorch Hub*[4]). Alternatively, platforms

3 https://www.tensorflow.org/hub

4 https://pytorch.org/hub/

such as *Hugging Face*[5]or *Kaggle*[6] are channels to source models that can be easily imported into common deep-learning frameworks in *Python*. More recent models are often only available as a set of scripts and files to be downloaded manually – which poses a significant inconvenience to researchers depending on their programming training. The following three subchapters will explain our embedding decisions.

Video

All state-of-the-art models are 3-dimensional convolutional neural networks, which differ in their performance only slightly across different classification tasks and training datasets (e.g., Huddar et al., 2020; Pandeya & Lee, 2021; Shang et al., 2021). Nevertheless, ideally, all current models should be tested if computationally feasible. For this research, we decided on *3D Resnet*[7], a state-of-the-art model available via the *PyTroch Hub*. It is trained on the aforementioned *Kinetics 400* dataset and used in its pre-trained version without additional fine-tuning.

In line with the preprocessing requirements of *3D Resnet*, we sampled 32 frames from each video equally distributed over the video's length[8]. Depending on the application, other sample techniques can be helpful. Scene detection algorithms can identify sufficiently distinct parts of a video or maybe only the first 2 seconds of a video are of interest because the user has only watched those (Qi et al., 2023; Tian et al., 2019).

The second preprocessing requirement of *3D Resent* is that the single frames need to have dimensions of 256*256 pixels. Therefore, we squished the frames to the desired format – compared to cropping, this preserves more visual information from the original frame – even in reduced granularity (see Figure 2). Cropping would need previous knowledge of the area within the videos to focus on – which we do not have in the case of TikTok posts.

For each video, the preprocessed 32 frames are then fed into the *3D Resent* model, and the last hidden layer (length = 2304) is retrieved as the feature representation for the respective TikTok video. The resulting feature vector has two dimensions (2304x32) representing an embedding for each of the video's input frames. To retrieve an embedding for the whole video, the 2D vector is reduced to a 1D vector through element-wise aggregation, such as averaging (Selva et al., 2023).

5 https://huggingface.co/

6 https://www.kaggle.com/

7 https://pytorch.org/hub/facebookresearch_pytorchvideo_resnet/

8 If a video is 16 seconds long and has 30 frames per second we sample every 15th frame.

Figure 2 Imapact of cropping versus squishing on one example frame. We can see that the squished frame retains, unlike the cropped frame, information on the green & orange pepper. Original photo modification of Flat-lay Photography of Variety of Vegetables [E. Akyurt]. (204), under a Creative Commons [0] license.

Audio

TikTok videos' audios are heterogeneous – voice, music, and action-related acoustic signals are all possible. To acknowledge this variety, *VGGish* is used. *VGGish* is developed by *Google LLC* and trained on the *AudioSet* database. *AudioSet* is based on 2.1 million *YouTube* videos trained on 527 classes, from music over speech to lawn mowing (Hershey et al., 2017).

Like the video embedding, we extracted the feature representation based on the last hidden layer of the model (length = 4096). The embeddings returned reflect each second of the input audio and are aggregated to a 1D vector via element-wise average aggregation.

Text

The video descriptions are multi-lingual. Investigating a subset of videos[9], we find predominantly German (38.51%) & English (30.16%) descriptions. But also Korean, Arabic, Turkish, Russian & Cantonese content (together 15%). The language detection was conducted with *fasttext* (Joulin et al., 2016). A content classifier should be able to handle multi-lingual data, given that we cannot control the language of content in the DDPs. We use a state-of-the-art multi-lingual BERT model (Reimers & Gurevych, 2019). The model *distiluse-base-multilingual-cased-v1* is used since it supports 15 languages, including all mentioned above except

9 The training dataset described later in this paper (N = 5,619).

Cantonese (1.5% of the descriptions). The output of the said model is not related to a classification taxonomy dictated via the training data but is supposed to serve as an input for further classification tasks. Therefore, we use only the final layer. *distiluse-base-multilingual-cased-v1*[10] returns a 1D vector of length 512.

After data collection and augmentation, each resulting aDDP (n = 18) consists of 1) the DDP, 2) corresponding survey data on sociodemographic characteristics and TikTok usage, 3) the raw content data (audio and video files that have been scraped), 4) metadata (length, likes, etc.), and 5) feature embeddings of the major modes a TikTok posts consists out of (visuals, audio & the textual description). All Python scripts used throughout the collection and augmentation process are made available open source (Wedel, 2024).

Applying aDDPs in TikTok Research

In the pilot study, we investigate the impact of user engagement behavior on the type of videos users encounter in their watch history. We use this exploratory question to showcase how aDDPs can be used in TikTok research and acknowledge that this is a proof-of-concept, not a study on its own. Results should, therefore, be interpreted accordingly. The user trace data under investigation are the 18 aDDPs, the collection and augmentation procedure of which has been described above.

As engagement behavior, we understand any action that signals a user paying attention to content. Here, we differentiate between passive (long watch time) and active (liking, sharing, etc.) engagement, along with the argumentation of first- and second-level exposure (Ohme & Mothes, 2020). The pilot study seeks to answer the following research questions concerning our 18 participants:

RQ1: Do users who show engagement behavior on informative videos receive more of such videos in future sessions/ within sessions?

RQ2: Does the users' self-reported consumption of informative videos align with actual digital trace data?

To facilitate research on the proposed questions, aDDPs are necessary because we need fine-grained user behavioral data (DDPs), survey data, and a database that allows us to classify each video with regard to whether it is informative or not (content data & multimodal feature representations).

However, DDPs do not let us know where the user has watched the videos on the platform. As of the time of data collection, TikTok holds two different feeds: the *for you feed* (algorithmically curated video suggestions) and the *following feed*

10 https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v1

(only videos published by creators a user follows). The *for you feed* is the default feed when opening the app and has been reported by TikTok as the dominant form of content consumption (TikTok, 2019). It is unclear to what extent suggestions from the following feed are also algorithmically suggested. Given that we can not distinguish between algorithmic and otherwise curated videos, we cannot certainly say that our results apply exclusively to the *for you feed*.

Methodology

To answer both research questions, we combine the different elements of the collected aDDPs. The DDPs were collected between the 18th of September 2023 and the 3rd of October from a German convenience sample, as described in the previous chapter on the TikTok DDP collection. The study sample comprised 18 individuals in Germany: 8 participants aged 16-26 and 10 aged 27-34. Most participants (15) held a university degree, while three did not. There were more females (9) than males (6), and three participants did not disclose their gender. The DDPs have been augmented as described in the respective previous section.

To facilitate content classification based on the multimodal feature embeddings, we train a classifier that categorizes the videos in the aDDPs into "informative" and "other" categories. The two categories are derived from the TikTok explore page classification. The TikTok explore page[11] is a website accessible via the TikTok desktop web interface. At the time of data collection, it consisted of 11 categories, where up to 200 videos were sorted within each category. The videos change constantly; to increase the dataset, we scraped the page repeatedly. The other ten categories are in contrast: *Dance & Music, Sports, Entertainment, Comedy & Drama, Cars, Fashion, Lifestyle, Pets & Nature, Relationships and Society.* The dataset is made available open source (Wedel, 2023). TikTok does not provide a description of these categories. A screening of the videos sorted under *Informative* shows mostly videos with tech, language, or finance tips and videos explaining scientific findings or history. We rely on the categorization being coherent enough to serve as a robust classification base for this proof-of-concept example. The chosen classification serves as an example of a prelabeled dataset that research needs to gather – either by manual labeling or using the limited number of videos labeled by TikTok.

The following sub-section guides through 1) the engagement measures based on the digital trace data, 2) the self-report-based engagement measures, and 3) the classifier training, including the subsequent classification of the videos in the aDDPs. We answer our RQs with binomial linear regression and the Pearson correlation coefficient.

11 https://www.tiktok.com/explore

The aDDP-based engagement measures

Each aDDP is split into sessions using the time stamps included in the DDP. Each session represents a user's consecutive consumption of videos without a break. The information within TikTok DDPs does not allow us to decide on the sessions with absolute certainty. To detect the session breakpoints, we use a threshold of 105 seconds that Zannettou et al. (2023) derived from 347 TikTok DDPs. That means that when there is an activity duration of more than 105 seconds, we count that as a breakpoint between two sessions of consecutive content consumption.

We operationalize passive engagement with a user having watched a video longer than their median watch time of a video. The watch time has been derived following previous studies via the timestamps for each video, and the last video in each session was removed from the dataset after deriving the watch time of the preceding video (Goetzen et al., 2023; Zannettou et al., 2023). Active engagement behavior encompasses all active actions that can be taken by a user concerning a video: liking, sharing, commenting, and favoring. For the sake of simplicity, we aggregate those actions as active engagements but acknowledge that this step depends on the research question – a more granular analysis is possible should the research question desire this.

During the preprocessing of the DDPs on the participant's local devices, an unstable sorting algorithm was used, which does not allow the above-described analysis for sessions with duplicate timestamps. Regarding two activities with the same timestamp, we do not know which came first. Therefore, it is impossible to know which video has been watched for x seconds, which has been directly skipped, or to which video a follow-up engagement action relates. Therefore, we excluded all sessions with duplicate timestamps from the analysis. This renders 47.45% (n = 12,750) of the overall detected sessions with more than one activity unusable, leaving 14,117 sessions for analysis. The exclusion of those sessions does not allow for empirical findings beyond within-session effects. Since the present study is meant to be solely a proof-of-concept, we nevertheless exemplarily measure cross-session effects.

Self-reported information exposure measures

To measure the participants' self-perception of information consumption, we asked participants to assess on a 5-point Likert scale how much they agreed (1 strongly disagree to 5 strongly agree) with the following four statements: a) *TikTok is important for me to stay up to date with current affairs (politics, economics, etc.). (M = 2.44, SD = 1.34) ; b) TikTok is important for me to stay up to date with general affairs (celebrities, sports, etc.). (M = 3.167, SD = 1.38)* ; c) *TikTok is important for me to learn new things (DIY, cooking, etc.). (M = 3.61, SD = 1.09) ; and d) TikTok is showing me primarily informative content (M = 2.344, SD = 1.15).*

The statements are based on past research on news use of young German adults on social media and cover the broader news categories of hard news (cur-

rent affairs) and soft news (general affairs) and summarize the remaining content[12] under learning and general information (Anter & Kümpel, 2023).

Training a classifier for aDDPs

To retrieve a pre-labeled dataset for model training, we scraped all videos on the TikTok explore page mentioned earlier from the 31st of July 2023 until the 4th of August 2023. While we retrieved around 200 unique videos per day – removing duplicates that occurred through videos being listed under one category for several days – due to the several dates of data collection, the initial training data set consisted of 473 videos labeled as informative and 4,664 videos tagged as a different category. An overview of the video overlap throughout the five days of scraping can be found in Appendix I.

To ease the unbalanced nature of the data, we decided to add the informative labeled videos from an earlier data collection (on the 4th, 12th, 13th, and 17th of July), resulting in 955 informative videos in total. Given the overall diversity of included categories, this training dataset of 5,619 unique videos can be assumed to represent a higher variation of videos compared to, e.g., keyword sampling methods that only include an often smaller number of videos from one specific domain while holding a meaningful number of instances of the target class. The metadata collection was facilitated via the *4CAT Toolkit (Peeters & Hagen, 2022)* and the *Zeeschumier (Peeters, 2023)* browser extension.

For classification, we tested a Support Vector Machine (SVM) as a traditional classifier for binary classification and a simple, fully connected Neural Network (NN) architecture with six hidden layers (see Appendix II). The target variable was the binary classification decision between *informative* and *other*. As model inputs, we tested uni- and multimodal representations based on the retrieved feature embeddings for three modalities of a video post (video, audio, text).

The critical design choice of a multimodal classifier is its fusion-mechanic (Sleeman et al., 2021). Fusion describes how the different modes are fused into one multimodal representation before (*early fusion*), during (*intermediate fusion*), or after (*late fusion*) the classification. For the case of TikTok, *early fusion* is sufficient since we can expect all modalities to be present (Choi & Lee, 2019). In early fusion, we concatenate the three calculated embeddings before we feed them into the tested classifiers to one embedding vector (e.g., multimodal representation of the respective video). Besides being easily implemented, early fusion also affords without effort the exploitation of cross-modality correlations (Zeppelzauer & Schopfhauser, 2016).

For the neural networks, each fully connected layer is followed by a dropout layer to avoid co-adaption within the network (Hinton et al., 2012). The hyperpa-

12 Tips and inspirations; Service; Consumption and welfare; Trivia, Activism; Comedy and fun

rameters for all neural networks were set at 50 epochs, a batch size of 40, a learning rate of 0.0001, and a dropout chance of 0.2 after hyperparameter tuning.

For both types of classifiers, we oversampled the minority class (informative) during training to be represented equally often compared to the majority class (other). Min-max normalization has been applied to each single-mode embedding vector based on all respective embeddings from the test, train, and inference corpus. Training and validation have been facilitated via 5-fold cross-validation. We report the average results over all five folds for precision, recall, F1-Score, and accuracy. The results show that all tested supervised machine learning techniques perform better than uniform random guessing – validating that all models pick up decisive features within the data to outperform an uninformed classification (see Table 2).

The common characteristic of the best-performing models (SVM^{T+A+V}, NN^{T+A+V}, SVM^{T}, NN^{T+A}) is that they include the text mode. The best-performing model is the single-mode SVM^{T}, closely followed by the trimodal models SVM^{T+A+V} and NN^{T+A+V}. The predictions of the NN^{T+A+V} and the SVM^{T} have a high variation in performance across the folds compared to the SVM^{T+A+V} (see Figure 3).

The SVM^{T} model classifies, on average, across all five-folds, 86.9% of the actual informative videos correctly, and 85.9% of the informative classified videos are indeed informative. Other tested models afford a higher recall, but the trade-off in terms of reduced precision always results in an overall reduced F1 score (see Table 2).

The learned classification is based on a classification by TikTok, and we are likely to reproduce an algorithmic error (see Boeschoten et al., 2020) that is part of TikTok's classification. Hence, future research needs to conduct robust (manual) training and validation data labeling. Tested models should be validated on a labeled sample from aDDPs videos to assess a model's performance appropriately on the set of videos in the aDDPs. Based on the used test and training data, this work shows that unimodal SVMs might be sufficient depending on the classification scheme and underlying data. Nevertheless, the within NN comparison also indicates that for NN classifiers, multimodality improves the classification significantly – supporting the assumption that they can exploit correlations between the modes. Given the recall and precision measures of the SVM^{T} model, we can assume that it misses ~15% of informative videos and misclassifies ~15% of them as "informative".

Table 2 Performance comparison of the tested classifiers. Maximum in bold. The first data row represents no trained model but shows the performance of uniform random guessing as a baseline. Performance values are reported, with "informative" being the target class. The model name reflects the classifier type (SVM = Support Vector Machine; NN = Neural Network) and the included modalities (T = Text; A = Audio; V = Video).

Modalities used	Model name (Base Classifier$^{\text{Initials of the modalities}}$)	Recall	Precision	F1-Score	Accuracy
-	Uniform random guess	.5	.11	.18	.493
Text	SVM^{T}	.869	.859	.864	.953
	NN^{T}	.881	.635	.730	.887
Audio	SVM^{A}	.751	.380	.505	.749
	NN^{A}	.781	.738	.758	.915
Video	SVM^{V}	.775	.597	.674	.873
	NN^{V}	.778	.819	.797	.933
Text + Audio	SVM^{T+A}	.915	.412	.568	.763
	NN^{T+A}	.909	.695	.781	.909
Text + Video	SVM^{T+V}	.813	.720	.763	.914
	NN^{T+V}	.916	.793	.844	.939
Video + Audio	SVM^{V+A}	.839	.801	.819	.937
	NN^{V+A}	.817	.814	.815	.937
Text + Audio + Video	SVM^{T+A+V}	.910	.804	.853	.947
	NN^{T+A+V}	.854	.856	.852	.949

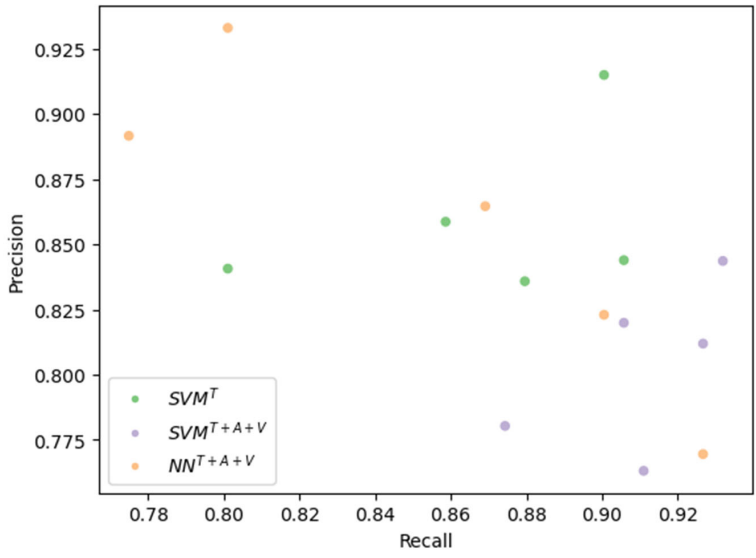


Figure 3 Precision and recall for each fold of the three best models by F1-score: the text only SVM (SVM^{T}), the trimodal support vector machine (SVM^{T+A+V}) and the trimodal neural network (NN^{T+A+V}).

Analysis & Results

Research question 1 asked if users who show engagement behavior on informative videos receive more of such videos in future sessions/ within sessions. For the present example, we first investigated the data on an aggregated level descriptively (see Table 3). We included the most recent 100 sessions, if available for each user, resulting in 11,475 videos over 1,242 sessions in our analysis. The SVM^{T} model labels 542 posts as "informative" and 10,933 as "other". The informative labeled videos make up, on average, 5% of the videos watched by our participants. The average profits here from one outlier – user 17, with 12% of their videos being informative. Regarding engagement behavior, our participants clearly show less engagement behavior (active and passive) towards the informative content in their feeds than the engagement behavior towards other content (2% vs. 47% for passive and 0% vs. 4% for active). We conclude that passive engagement behavior for videos labeled as informative and active engagement behavior, in general, is sparse among our participants.

Table 3 Fraction of informative videos and engagement behavior aggregated per user.

user	informative videos	passive engagement		active engagement		#sessions	#videos
		info	other	info	other		
1	.03	.01	.18	.00	.02	100	1482
2	.02	.01	.66	.00	.04	31	140
3	.05	.01	.05	.00	.00	100	2002
4	.05	.03	.67	.00	.09	57	286
5	.07	.03	.43	.01	.07	97	674
6	.04	.02	.62	.00	.01	94	642
7	.06	.05	.75	.00	.01	16	93
8	.00	.00	.90	.00	.00	2	10
9	.08	.02	.70	.00	.01	23	120
10	.03	.02	.71	.01	.40	83	373
11	.02	.00	.09	.00	.00	100	1349
12	.03	.00	.20	.00	.00	96	1050
13	.05	.02	.54	.00	.01	97	650
14	.06	.02	.56	.00	.02	34	349
15	.05	.02	.40	.00	.05	98	733
16	.03	.00	.36	.00	.00	88	452
17	.12	.01	.13	.00	.01	58	911
18	.07	.03	.54	.00	.00	9	159
mean	.05	.02	.47	0	.04	65.72	637.50
total	-	-	-	-	-	461	3626

We then applied a binomial regression model with the respective engagement behaviors as an independent variable (IV) and the fraction of informative videos as the dependent variable (DV). We investigate two possible correlations: First, across sessions, the IV represents the fraction of engagement behavior on informative videos in sessions s_t, and the DV represents the fraction of informative videos in the following sessions s_{t+1}. Second, within sessions, the IV represents the fraction of engagement behavior on informative videos in the first half of session s, and the DV represents the fraction of informative videos in the second half of session s.

Given the sparsity of engagement behavior, we could not analyze all users for both engagement behaviors. For the users where the analysis could be conducted, we find no implications that the IV and DV correlate with three exceptions (user 4 and 6 passive and user 15 active) (see Table 4). This means that only in three cases is there an indication of a relationship between previous usage behaviors and the amount of future informative videos, suggesting that engagement behavior on a specific type of video will lead to users having more of such videos in their following sessions. RQ1, hence, cannot be answered affirmatively. Moreover, future research would need to apply time series analysis to investigate the causal direction of the relationship and, the time lag between engagements and possible effects on suggested videos.

Table 4 Binomial regression results.

		Across sessions		Within sessions	
		p	r-squared	p	r-squared
1	passive	.8	0.025765	.671	0.060967
	active	.837	-0.020926	.248	0.1648
2	passive	.481	-0.133834	-	-
	active	.636	0.089981	-	-
3	passive	.731	0.034955	.715	-0.050793
4	passive	.041*	0.274038	.626	-0.255133
	active	.479	0.096461	.165	0.647343
5	passive	.258	0.116413	.843	-0.041792
	active	.096	0.170661	.799	0.053595
6	passive	.019*	0.241892	.683	0.094799
	active	.969	0.004048	.455	0.172234
7	passive	.537	-0.173152	-	-
	active	.8	0.071429	-	-
8	No sufficient engagement data on informative content.				
9	passive	.158	0.311532	-	-
	active	.598	0.119063	-	-
10	passive	.417	-0.090855	.527	0.32659
	active	.185	-0.147704	.792	-0.139792
11	passive	.198	0.130589	.803	0.038768
	active	.572	0.057455	-	-
12	passive	.225	-0.12557	.748	-0.050957
	active	.558	0.060883	.573	0.089542
13	passive	.698	-0.040057	.845	0.046675
	active	.942	-0.007459	.384	-0.205696

		Across sessions		Within sessions	
		p	r-squared	p	r-squared
14	passive	.285	-0.191583	.128	-0.545731
	active	.246	-0.207755	.066	-0.634986
15	passive	.547	-0.061846	.361	0.210043
	active	.375	0.090977	.038*	0.455085
16	passive	.384	-0.094379	.427	-0.267155
17	passive	.522	0.086542	.886	-0.028408
	active	.425	-0.107593	.499	0.133362
18	passive	.133	0.579	.944	0.088475

Given the methodological nature of this paper, the analysis should not be taken as empirical evidence. The respective methodological pipeline is not grounded on a robust definition of *informative*. Nevertheless, with regards to the TikTok-defined term of informative videos for the majority of the participants, we do not find their engagement behavior impacting the fraction of informative content - neither within sessions - nor across sessions. The results for the cross-session comparison are unreliable, given the number of sessions that had to be excluded for the analysis because of duplicate timestamps.

Research question 2 asked for the relationship between self-reported content consumption and actual consumption of informative content on TikTok. Here, the full breadth of an augmented DDP can be used, as we rely on the survey data gathered from participants. Based on the self-reported data and the multimodal classification of the videos in a user watch list, we can test how closely users' self-perception comes to their digital behavior.

Self-reported information consumption was collected for *current affairs*, *general affairs*, *learning,* and *general information*. We again used the fraction of informative videos within each participant's 100 most recent sessions for the observed behavior. Analysis revealed a negligible correlation for *current affairs* ($r = 0.268$, $p = 0.282$), *learning* ($r = 0.226$, $p = 0.366$), and *general information* ($r = -0.178$, $p = 0.478$), a moderate correlation has been found for *general affairs* ($r = 0.507$, $p = 0.032$). Previous research (e.g., Araujo et al., 2017; Ohme et al., 2021; Parry et al., 2021) has shown that users' self-reports deviate from the observed digital behavior. Our pilot study suggests similar patterns for all dimensions of informative other than general affairs.

We note that we asked for qualitative assessments ("How much do you agree with …"), not for quantitative ("How often do you consume …") in terms of content consumption. The question items are less comparable with the cited studies – given that those explicitly asked for a quantification of content consumption.

Discussion

aDDPs present a promising future for digital trace data analysis. With open-source tools such as *Port* (Boeschoten et al., 2023), the means to collect such data is accessible to the broader research community. Using such tools also increases the transparency of the data collection. aDDPs are non-reactive and thereby come without the caveats of data collection methods that can compete otherwise (partially) with the collected data's granularity (e.g., tracking apps) or its modality (e.g., screenshot apps). The combination of granular information about user traces and the richness of publicly available video content data assessed through the initial DDPs make aDDPs an unprecedented database for critical social science research.

The paper presents a systematic approach to augmenting DDPs with multimodal data and using such data to answer substantial research questions. We do that specifically for TikTok, but this approach is flexible and adaptable to other data download packages. Augmenting DDPs of a multimodal nature presents a challenge to current research and has not been done before. This paper presents a unique approach with a clear pipeline on how to proceed with such an endeavor. It is a proof-of-concept on how content features of TikTok videos can be included in social science research, sampled via data donations.

Right now, aDDPs are especially helpful for vertical video platforms (VVPs) because researchers can collect the watch history retrospectively for half a year. The limit of half a year in the case of TikTok is a notable restriction, in line with the general unreliability and volatility of DDPs from different platforms (Carriére, 2023). It is not transparent whether the limitation comes from TikTok not saving the watch history for a user longer than half a year or if they only provide limited data.[13] Therefore, the *Digital Service Act* is a welcome prospect for improving the conditions for scientific work on user trace data – implementing an infrastructure that enforces transparency and scientific data access (Hase et al., 2023).

For TikTok DDPs, specifically session detection and the question of how a post was encountered (through the *for you feed* or else) are unsolved methodological questions. Here, it is similarly desirable that the platforms deliver even more detailed trace data. To detect session breakpoints, one could use the login timestamps. An initial attempt showed that they do not consistently mark the beginning of a session – users might stay logged in for a session break. While being reliable, login timestamps are not entirely sufficient.

Regarding the collection of DDPs, we must stay attentive to the difficulties and biases. Out of the 42 people who opened the survey invitation, only 18 donated their data. Future studies must carefully consider the reasons for the willing-

13 The suspicion originates especially from other activates such as following and liking being part of the DDP for the whole duration of the accounts existence.

ness to donate data for the platform of their interest (e.g., Pfiffner & Friemel, 2023). It remains a discussion within data donation studies to what degree classic representative samples are achievable. Nevertheless, for many research questions, answers coming from an in-depth analysis of online behavior coming from the digital traces of a specific subgroup may be a welcome complement to results from representative samples that are only able to rely on self-reports.

For 13.58 % of the analyzed subset of videos found in the data donations, we could not retrieve any metadata anymore and, thereby, for a similar fraction of videos as in previous studies on TikTok DDPs (Zannettou et al., 2023). Digital trace data from TikTok has the same limitations as trace data from other platforms. For the reproducibility of subsequent research, only the unique identifier of a video should be shared, not the content itself, to ensure the right to be forgotten on the video creator's side (General Data Protection Regulation, Regulation (EU) 2016/679, Art. 17; 2016). As conducted for other social media platforms, systematic research on the impact of no longer available content for TikTok is needed (e.g., Buehling, 2023; Zubiaga, 2018):

This paper is one of the first to compare uni- and multimodal classifications of TikTok videos, traditional machine learning, and deep learning approaches. Yet, we acknowledge that the classification is roughly cut, and more relevant content categories will need a robust definition on which basis a training and validation data set is manually labeled. Given the breadth of variation that multimodal representation with thousands of features proposes, we estimate that a minimum of 1000 videos for each class is desirable. However, further research is needed to explore the actual sample sizes.

The classification models have shown that an unimodal traditional machine-learning approach was sufficient. Looking only at the neural networks shows that the trimodal neural network performs the best. Neural networks hold a high potential for improvement. Optimizations like a more sophisticated architecture (e.g., Shang et al., 2021; Tian et al., 2019) or better input data can lead to them superseding traditional machine learning classifiers for multimodal classification tasks. A juvenile indicator for that is that except for the text-only models, the neural network-based models performed generally better or were similarly suitable for all other test conditions.

We must also acknowledge that augmenting data introduces errors in the observed data. While self-reported user measures suffer from recall biases, augmented DDPs suffer from algorithmic errors that are an irreducible part of the pre-trained models employed to retrieve embeddings for each modality and missing data errors through DDPs only covering a fraction of an individual's media environment. We need to be aware that despite the great future of digital trace data, getting closer to a ground truth may be possible, but reaching it will remain a challenge.

While we showcase here the usefulness of an aDDP and the possibilities for substantial research, errors can be introduced in each part of the data collection, augmentation, and analysis. Future research should, therefore, apply the total error framework (Boeschoeten et al. 2022) when preparing an aDDP.

Augmentation needs resources, both from a human and a computational perspective. Doing this for a single research process is challenging, and we suggest working in greater collaborations, whereas 'seed DDPs' can be increasingly augmented - growing over time. Such consortiums could work together in larger data collection cycles to reach more significant and more complex datasets to answer multiple research questions (e.g., Ohme et al., 2023) or – assuming adequate privacy, ethical, and security measures – combine DDPs from different data collection cycles and automatically augment them. There remains a discussion as to under which conditions and how the data collected can be reused and shared – which would drastically increase the accessibility of the method to researchers unable to scrape or code. Examining this against EU, national, and institutional regulations would be the priority of such a consortium.

This study has shown that aDDPs open up new spheres of research. With such a procedure, researchers are not merely bound to the information the donations carry but can investigate a plethora of questions that rely on classifications that the platforms do not provide. aDDPs unite user-centric and content data collection. Embracing an aDDP allows research to expand questions on the distribution of anti-vax (Kim et al., 2023) or sexualized content (George & Surdeanu, 2023) with a user-centered perspective: *What do users actually see, and how do they react to it?* Vice-versa, do aDDPs allow studies that focus on user-centric data (e.g., survey, data donations) to cover more depth instead of relying purely on an existing data basis for the classifications of actors or domains or solely on the available metadata (Zannettou et al., 2023):

In a time when visual online platforms such as TikTok, YouTube Shorts, or Instagram have grown more prevalent – and with them an entirely new level of reliance on visual cues instead of textual description – it is as relevant as ever to explore the means to analyze such online content. Be it to explore the algorithmic curation of those new platforms, the harm they might do, or their impact on opinion formation. Consequently, this paper introduces a novel methodological framework to enhance the study of visual online platforms, enabling social science researchers to address previously inaccessible research questions.

Bibliography

Akyurt, E. (2024). *Flat-lay Photogtaphy of Variety of Vegtables*. Pexels. https://www.pexels.com/photo/flat-lay-photography-of-variety-of-vegetables-1435904/

Anter, L., & Kümpel, A. S. (2023). Young Adults' Information Needs, Use, and Understanding in the Context of Instagram: A Multi-Method Study. *Digital Journalism*, *0*(0), 1–19. https://doi.org/10.1080/21670811.2023.2211635

Araujo, T., Ausloos, J., Atteveldt, W. van, Loecherbach, F., Moeller, J., Ohme, J., Trilling, D., Velde, B. van de, Vreese, C. de, & Welbers, K. (2022). OSD2F: An Open-Source Data Donation Framework. *Computational Communication Research*, *4*(2), 372–387. https://doi.org/10.5117/CCR2022.2.001.ARAU

Araujo, T., Wonneberger, A., Neijens, P., & de Vreese, C. (2017). How Much Time Do You Spend Online? Understanding and Improving the Accuracy of Self-Reported Measures of Internet Use. *Communication Methods and Measures*, *11*(3), 173–190. https://doi.org/10.1080/19312458.2017.1317337

Baltrusaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(2), 423–443. https://doi.org/10.1109/tpami.2018.2798607

Basch, C. H., Hillyer, Grace C., & Jaime, Chistie. (2020). *COVID-19 on TikTok: Harnessing an emerging social media platform to convey important public health messages*. https://doi.org/10.1515/ijamh-2020-0111

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Boeschoten, L., Ausloos, J., Moeller, J., Araujo, T., & Oberski, D. L. (2020). *A framework for digital trace data collection through data donation* (No. arXiv:2011.09851). arXiv. https://doi.org/10.48550/arXiv.2011.09851

Boeschoten, L., Schipper, N. C. de, Mendrik, A. M., Veen, E. van der, Struminskaya, B., Janssen, H., & Araujo, T. (2023). Port: A software tool for digital data donation. *Journal of Open Source Software*, *8*(90), 5596. https://doi.org/10.21105/joss.05596

Boeschoten, L., Voorvaart, R., Van Den Goorbergh, R., Kaandorp, C., & De Vos, M. (2021). Automatic de-identification of data download packages. *Data Science*, *4*(2), 101–120. https://doi.org/10.3233/DS-210035

Buehling, K. (2023). Message Deletion on Telegram: Affected Data Types and Implications for Computational Analysis. *Communication Methods and Measures*, *0*(0), 1–23. https://doi.org/10.1080/19312458.2023.2183188

Carriére, T. (2023, December 9). *Volatility of Data Download Packages*. Data Donation Symposium, Zurich. https://datadonation.uzh.ch/en/symposium-2023/

Chiatti, A., Davaasuren, D., Ram, N., Mitra, P., Reeves, B., & Robinson, T. (2019). *Guess What's on my Screen? Clustering Smartphone Screenshots with Active Learning*. http://arxiv.org/pdf/1901.02701v2

Choi, J.-H., & Lee, J.-S. (2019). EmbraceNet: A robust deep learning architecture for multimodal classification. *Information Fusion*, *51*, 259–270. https://doi.org/10.1016/j.inffus.2019.02.010

DataSkop. (2023). *Wie tickt TikTok?* DataSkop. https://dataskop.net

Driel, I. I. van, Giachanou, A., Pouwels, J. L., Boeschoten, L., Beyens, I., & Valkenburg, P. M. (2022). Promises and Pitfalls of Social Media Data Donations. *Communication Methods and Measures*. https://doi.org/10.1080/19312458.2022.2109608

DSA: Very Large Online Platforms and Search Engines. (2023). [Text]. European Commission - European Commission. https://ec.europa.eu/commission/presscorner/detail/en/IP_23_2413

Egger, R., Kroner, M., & Stöckl, A. (2022). Web Scraping. In R. Egger (Ed.), *Applied Data Science in Tourism: Interdisciplinary Approaches, Methodologies, and Applications* (pp. 67–82). Springer International Publishing. https://doi.org/10.1007/978-3-030-88389-8_5

Fiallos, A., Fiallos, C., & Figueroa, S. (2021). Tiktok and Education: Discovering Knowledge through Learning Videos. *2021 Eighth International Conference on eDemocracy & eGovernment (ICEDEG)*, 172–176. https://doi.org/10.1109/ICEDEG52154.2021.9530988

Freelon, D. (2023). *Dfreelon/pyktok* [Python]. https://github.com/dfreelon/pyktok (Original work published 2022)

Friemel, T. N., & Pfiffner, N. (2023). *The Data Donation Module*. https://datadonation.uzh.ch/en/infrastructure/

General Data Protection Regulation, Regulation (EU) 2016/679, Art. 17, § Uncategorized (2016). https://gdpr.eu/article-17-right-to-be-forgotten/

George, E., & Surdeanu, M. (2023). *It is not Sexually Suggestive, It is Educative. Separating Sex Education from Suggestive Content on TikTok Videos* (No. arXiv:2307.03274). arXiv. https://doi.org/10.48550/arXiv.2307.03274

Goetzen, A., Wang, R., Redmiles, E. M., Zannettou, S., & Ayalon, O. (2023). *Likes and Fragments: Examining Perceptions of Time Spent on TikTok* (No. arXiv:2303.02041). arXiv. https://doi.org/10.48550/arXiv.2303.02041

Haim, M., Leiner, D., & Hase, V. (2023). Integrating Data Donations into Online Surveys. *Software Review*.

Hase, V. (2023, November 9). *Fulfilling data access obligations: Platforms need to increase their compliance to enable data donation studies*. Data Donation Symposium, Zurich. https://datadonation.uzh.ch/en/symposium-2023/

Hase, V., Boczek, K., & Scharkow, M. (2022). Adapting to Affordances and Audiences? A Cross-Platform, Multi-Modal Analysis of the Platformization of News on Facebook, Instagram, TikTok, and Twitter. *Digital Journalism*. https://doi.org/10.1088/1751-8113/42/3/035201

Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., Slaney, M., Weiss, R. J., & Wilson, K. (2017). *CNN Architectures for Large-Scale Audio Classification* (No. arXiv:1609.09430). arXiv. https://doi.org/10.48550/arXiv.1609.09430

Hiippala, T. (2017). The Multimodality of Digital Longform Journalism. *Digital Journalism*, *5*(4), 420–442. https://doi.org/10.1080/21670811.2016.1169197

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. (2012). *Improving neural networks by preventing co-adaptation of feature detectors* (No. arXiv:1207.0580). arXiv. http://arxiv.org/abs/1207.0580

Huddar, M. G., Sannakki, S. S., & Rajpurohit, V. S. (2020). Multi-level feature optimization and multimodal contextual fusion for sentiment analysis and emotion classification. *Computational Intelligence*, *36*(2), 861–881. https://doi.org/10.1111/coin.12274

Ibañez, M., Sapinit, R., Reyes, L. A., Hussien, M., Imperial, J. M., & Rodriguez, R. (2021). Audio-Based Hate Speech Classification from Online Short-Form Videos. *2021 International Conference on Asian Language Processing (IALP)*, 72–77. https://doi.org/10.1109/IALP54817.2021.9675250

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). *Bag of Tricks for Efficient Text Classification* (No. arXiv:1607.01759). arXiv. https://doi.org/10.48550/arXiv.1607.01759

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., & Zisserman, A. (2017). *The Kinetics Human Action Video Dataset* (No. arXiv:1705.06950). arXiv. https://doi.org/10.48550/arXiv.1705.06950

Kim, S. J., Villanueva, I. I., & Chen, K. (2023). Going Beyond Affective Polarization: How Emotions and Identities are Used in Anti-Vaccination TikTok Videos. *Political Communication*, *0*(0), 1–20. https://doi.org/10.1080/10584609.2023.2243852

Lepa, S., & Suphan, A. (2019). *Der Elefant im Wohnzimmer der Kommunikationswissenschaft: Die rechnergestützte Analyse nonverbaler digitaler Kommunikation*. https://doi.org/10.25598/JKM/2019-10.6

Li, L., & Kang, K. (2023). *Exploring the Relationships between Cultural Content and Viewers' Watching Interest: A Study of Tiktok Videos Produced by Chinese Ethnic Minority Groups*. 37–46. https://www.scitepress.org/Link.aspx?doi=10.5220/0010610900370046

Meßmer, A.-K., Degeling, M., & Jaursch. (2023). *Response to the European Commission's call for evidence on a planned Delegated Regulation on data access provided for in the Digital Services Act (DSA)*. Stiftung Neue Verantowortung.

Ming, S., Han, J., Li, M., Liu, Y., Xie, K., & Lei, B. (2023). TikTok and adolescent vision health: Content and information quality assessment of the top short videos related to myopia. *Frontiers in Public Health*, *10*. https://www.frontiersin.org/articles/10.3389/fpubh.2022.1068582

Mordecai, C. (2023). #anxiety: A multimodal discourse analysis of narrations of anxiety on TikTok. *Computers and Composition*, *67*, 102763. https://doi.org/10.1016/j.compcom.2023.102763

Newman, N., Fletcher, R., Eddy, K., Robertson, C. T., & Nielsen, R. K. (2023). *Reuters Institute Digital News Report 2023*.

Ng, R., & Indran, N. (2023). Videos about older adults on TikTok. *PLOS ONE*, *18*(8), e0285987. https://doi.org/10.1371/journal.pone.0285987

Ohme, J., Araujo, T., Boeschoten, L., Freelon, D., Ram, N., Reeves, B., & Robinson, T. N. (2023). *Digital Trace Data Collection for Social Media Effects Research: APIs, Data Donation, and (Screen) Tracking*. https://www.tandfonline.com/doi/full/10.1080/19312458.2023.2181319

Ohme, J., Araujo, T., Vreese, C. H., & Piotrowski, J. T. (2021). Mobile data donations: Assessing self-report accuracy and sample biases with the iOS Screen Time function. *Mobile Media & Communication*, *9*(2), 293–313. https://doi.org/10.1177/2050157920959106

Ohme, J., & Mothes, C. (2020). What Affects First- and Second-Level Selective Exposure to Journalistic News? A Social Media Online Experiment. *Journalism Studies*, *21*(9), 1220–1242. https://doi.org/10.1080/1461670X.2020.1735490

Pandeya, Y. R., & Lee, J. (2021). Deep learning-based late fusion of multimodal information for emotion classification of music video. *Multimedia Tools and Applications*, *80*(2), 2887–2905. https://doi.org/10.1007/s11042-020-08836-3

Parry, D. A., Davidson, B. I., Sewall, C. J. R., Fisher, J. T., Mieczkowski, H., & Quintana, D. S. (2021). A systematic review and meta-analysis of discrepancies between logged and self-reported digital media use. *Nature Human Behaviour, 5*(11), Article 11. https://doi.org/10.1038/s41562-021-01117-5

Peeters, S. (2023). *Zeeschuimer* [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.8399900

Peeters, S., & Hagen, S. (2022). The 4CAT Capture and Analysis Toolkit: A Modular Tool for Transparent and Traceable Social Media Research. *Computational Communication Research, 4*(2), 571–589. https://doi.org/10.5117/CCR2022.2.007.HAGE

Pfiffner, N., & Friemel, Thomas. N. (2023). Leveraging Data Donations for Communication Research: Exploring Drivers Behind the Willingness to Donate. *Communication Methods and Measures, 17*(3), 227–249. https://doi.org/10.1080/19312458.2023.2176474

Pfiffner, N., Witlox, P., & Friemel, T. N. (2022). *Data Donation Module (Version 1.0.0)* [Computer software]. https://github.com/uzh/ddm

Primig, F., Szabó, H. D., & Lacasa, P. (2023). Remixing war: An analysis of the reimagination of the Russian–Ukraine war on TikTok. *Frontiers in Political Science, 5*. https://www.frontiersin.org/articles/10.3389/fpos.2023.1085149

Qi, P., Bu, Y., Cao, J., Ji, W., Shui, R., Xiao, J., Wang, D., & Chua, T.-S. (2023). FakeSV: A Multimodal Benchmark with Rich Social Context for Fake News Detection on Short Video Platforms. *Proceedings of the AAAI Conference on Artificial Intelligence, 37*(12), Article 12. https://doi.org/10.1609/aaai.v37i12.26689

Ram, N., Yang, X., Cho, M.-J., Brinberg, M., Muirhead, F., Reeves, B., & Robinson, T. N. (2020). Screenomics: A New Approach for Observing and Studying Individuals' Digital Lives. *Journal of Adolescent Research, 35*(1), 16–50. https://doi.org/10.1177/0743558419883362

Reeves, B., Ram, N., Robinson, T. N., Cummings, J. J., Giles, C. L., Pan, J., Chiatti, A., Cho, M. J., Roehrick, K., Yang, X., Gagneja, A., Brinberg, M., Muise, D., Lu, Y., Luo, M., Fitzgerald, A., & Yeykelis, L. (2021). Screenomics: A Framework to Capture and Analyze Personal Life Experiences and the Ways that Technology Shapes Them. *Human-Computer Interaction, 36*(2), 150–201. https://doi.org/10.1080/07370024.2019.1578652

Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks* (No. arXiv:1908.10084). arXiv. http://arxiv.org/abs/1908.10084

Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. *Communications of the ACM, 63*(12), 54–63. https://doi.org/10.1145/3381831

Selva, J., Johansen, A. S., Escalera, S., Nasrollahi, K., Moeslund, T. B., & Clapés, A. (2023). Video Transformers: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–20. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/TPAMI.2023.3243465

Shang, L., Kou, Z., Zhang, Y., & Wang, D. (2021). A Multimodal Misinformation Detector for COVID-19 Short Videos on TikTok. *2021 IEEE International Conference on Big Data (Big Data)*, 899–908. https://doi.org/10.1109/BigData52589.2021.9671928

Sleeman, W. C., Kapoor, R., & Ghosh, P. (2021). Multimodal Classification: Current Landscape, Taxonomy and Future Directions. *arXiv: Learning*.

Stier, S., Breuer, J., Siegers, P., & Thorson, K. (2020). Integrating Survey Data and Digital Trace Data: Key Issues in Developing an Emerging Field. *Social Science Computer Review, 38*(5), 503–516. https://doi.org/10.1177/0894439319843669

Syed, M. S. S., Pirogova, E., & Lech, M. (2021). Prediction of Public Trust in Politicians Using a Multimodal Fusion Approach. *Electronics, 10*(11), Article 11. https://doi.org/10.3390/electronics10111259

Teather, D. (2023). *TikTokAPI* (Version 6.1.1) [Python]. https://github.com/davidteather/tiktok-api (Original work published 2019)

Tian, H., Tao, Y., Pouyanfar, S., Chen, S.-C., & Shyu, M.-L. (2019). Multimodal deep representation learning for video classification. *World Wide Web, 22*. https://doi.org/10.1007/s11280-018-0548-3

TikTok. (2019, August 16). *How TikTok recommends videos #ForYou*. Newsroom | TikTok. https://newsroom.tiktok.com/en-us/how-tiktok-recommends-videos-for-you

TiKTok. (2023). *TikTok Research API Terms of Service*. https://www.tiktok.com/legal/page/global/terms-of-service-research-api/en

TikTok. (2023a). *TikTok's DSA Transparency Report 2023*.

TikTok. (2023b, January 30). *Requesting your data | TikTok Help Center*. https://support.tiktok.com/en/account-and-privacy/personalized-ads-and-data/requesting-your-data

Valkenburg, P. M. (2022). Theoretical Foundations of Social Media Uses and Effects. In *Handbook of Adolescent Digital Media Use and Mental Health* (pp. 39–60). Cambridge University Press. https://doi.org/10.1017/9781108976237.004

Wedel, L. (2023). *A categorized multimodal TikTok dataset*. https://www.weizenbaum-library.de/handle/id/420

Wedel, L. (2024). *Augmented TikTok Data Donation Packages Repository* (Version 0.1) [Computer software]. https://github.com/lionwedel/augmented_tiktok_DDP/blob/main/README.md

Yeung, A., Ng, E., & Abi-Jaoude, E. (2022). TikTok and Attention-Deficit/Hyperactivity Disorder: A Cross-Sectional Study of Social Media Content Quality. *The Canadian Journal of Psychiatry, 67*(12), 899–906. https://doi.org/10.1177/07067437221082854

Zannettou, S., Nemeth, O.-N., Ayalon, O., Goetzen, A., Gummadi, K. P., Redmiles, E. M., & Roesner, F. (2023). *Leveraging Rights of Data Subjects for Social Media Analysis: Studying TikTok via Data Donations* (No. arXiv:2301.04945). arXiv. https://doi.org/10.48550/arXiv.2301.04945

Zeppelzauer, M., & Schopfhauser, D. (2016). Multimodal classification of events in social media. *Image and Vision Computing, 53*, 45–56. https://doi.org/10.1016/j.imavis.2015.12.004

Zhang, H., & Peng, Y. (2022). Image Clustering: An Unsupervised Approach to Categorize Visual Data in Social Science Research. *Sociological Methods & Research*, 004912412210826. https://doi.org/10.1177/00491241221082603

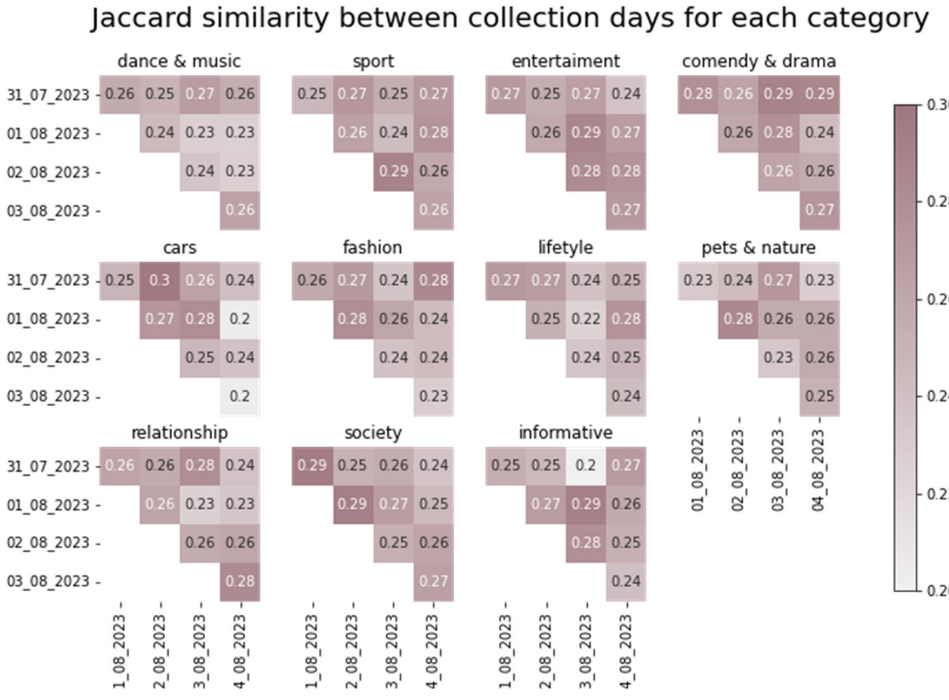
Zhou Ting. (2021). The Media Images of Old Influencers on TikTok: A Multimodal Critical Discourse Analysis. *Journal of Literature and Art Studies, 11*(10). https://doi.org/10.17265/2159-5836/2021.10.013

Zubiaga, A. (2018). A longitudinal assessment of the persistence of twitter datasets. *Journal of the Association for Information Science and Technology, 69*(8), 974–984. https://doi.org/10.1002/asi.24026

Zulko. (2023). *MoviePy* [Python]. https://github.com/Zulko/moviepy (Original work published 2013)

Appendix

I – Overlap measured by Jaccard similarity in unique videos between the five consecutive days of data collection for each category.



II – Neural Network Architecture

We used a Neural Network with six fully connected layers, with ReLu activation functions and five dropout layers for all input combinations. Below, we report the architecture as constructed in *PyTorch*. The layer size varies depending on the size of the input vector (number of input modalities). These in- and out-feature sizes adapted accordingly and always aimed to give the network a funnel shape.

```
six_layer(
  (classifier): Sequential(
    (0): Linear(in_features=6912, out_features=4096, bias=True)
    (1): ReLU(inplace=True)
    (2): Dropout(p=0.2, inplace=False)
    (3): Linear(in_features=4096, out_features=2048, bias=True)
    (4): ReLU(inplace=True)
    (5): Dropout(p=0.2, inplace=False)
    (6): Linear(in_features=2048, out_features=1024, bias=True)
    (7): ReLU(inplace=True)
    (8): Dropout(p=0.2, inplace=False)
    (9): Linear(in_features=1024, out_features=512, bias=True)
    (10): ReLU(inplace=True)
    (11): Dropout(p=0.2, inplace=False)
    (12): Linear(in_features=512, out_features=256, bias=True)
    (13): ReLU(inplace=True)
    (14): Dropout(p=0.2, inplace=False)
    (15): Linear(in_features=256, out_features=1, bias=True)
  )
  (sigmoid): Sigmoid()
)
```

Online first publication date: 15-10-2024