News on TikTok: An Annotated Dataset of TikTok Videos from German-Speaking News Outlets in 2023

Anna-Theresa Mayer[1], Lion Wedel[1], Jan Batzner[1], Jonathan Hendrickx[2], Emma Bremer[1], Alexander Iwan[1], Volker Stocker[1], Jakob Ohme[1]

[1] Weizenbaum Institute, Berlin, Germany
[2] University of Copenhagen, Denmark
anna-theresa.mayer@weizenbaum-institut.de

Abstract

TikTok has emerged as a leading social media platform with increasing relevance for the consumption and distribution of news, especially for younger age groups. Despite its growing relevance, analyses of how traditional news outlets produce content for the platform and maintain journalistic news values are limited. Moreover, there are few large-scale datasets that are suitable for tracing larger journalistic trends and developing approaches for the automated annotation of multimodal content. This paper addresses these gaps and introduces "News on TikTok," an annotated dataset of 8,623 TikTok videos published by 18 major German-speaking news outlets in 2023. Combining metadata with human-annotated data of 25 variables (incl. the presence of visual, auditory, and interactive elements, and journalistic news values), our dataset makes three significant contributions: First, it enables extensive analyses of news characteristics on TikTok. Second, it provides ground truth data to develop and validate automated tools for multimodal content analyses. Third, it offers a comprehensive guide for generating datasets with similar research interests.

Dataset — https://doi.org/10.7802/2863

Introduction

Over the last few years, TikTok has developed into a leading global social media platform, shifting from pushing lip-syncing content to becoming a key source of news and political information. The Reuters Institute Digital News Report 2024 confirms this growth, stating that "the proportion using it for news has grown to 13% (+2) across all markets and 23% for 18–24s" (Newman et al. 2024, p.12). Similarly, a Pew Research Center analysis shows that the share of US adults who regularly use TikTok to get news there has increased from 3% in 2020 to 17% in 2024 (Leppert and Matsa 2024). Unsurprisingly, the youngest age group (18-29-year-olds) has the highest preference, with numbers increasing from 9% in 2020 to 39% in 2024 (Leppert and Matsa 2024).

TikTok has also changed the digital media ecosystem through its algorithmic feed and short content formats. Evolving preferences for short video content (Newman et al. 2024, p.13), for example, require traditional news outlets to adapt how they produce and distribute news content. First studies have begun exploring news content on TikTok. However, current and comprehensive analyses of how traditional news outlets produce TikTok content and how they adhere to news values are limited. News values can be understood as criteria for the journalistic selection and production of news and include, e.g., positive and negative news, conflict, and scope (Galtung and Ruge 1965; Harcup and O'Neill 2017). Moreover, the growing volume of uploaded content and the expanded scope of potential communication modes within a single video (incl. text, image, and audio) necessitate automated and scalable approaches to analyze such multimodal content.

Motivated by these gaps, this paper introduces the dataset "News on TikTok." The dataset encompasses all annotated TikTok videos published in 2023 by 18 major German-speaking news outlets (N = 8,623). The dataset, available as open access, contains the videos' metadata obtained via the TikTok Research API and human-annotated data based on 25 variables in a supplementary codebook. Among our multidisciplinary team of authors, the first four iteratively developed the codebook to cover formal variables (incl. the visual, auditory, and interactive elements in the videos) and content variables regarding the general topic, news format, and the presence of news values in the videos.

The dataset "News on TikTok" makes three main contributions. First, it provides a large-scale dataset for the systematic exploration of news content on TikTok and offers novel opportunities to analyze the general characterization of news videos and the presence of news values, a concept central to journalism studies, on TikTok. Second, it provides human-annotated ground truth data that can serve as a benchmark to develop and validate automatized tools for analyzing visual and auditory elements of news content, by offering a database for the comparative evaluation of human and AI-based annotations in (computational) social science and beyond. Third, the dataset paper provides a comprehensive guide for creating datasets that can be used to explore similar research interests across different geographic and thematic contexts, or for comparative analysis.

Related Work

Previous studies have compared news outlets' content published on TikTok to the content they published on other social media platforms, such as Instagram, or on their respective websites, and have applied a single or cross-country scope for the selection of news outlets (Hase, Boczek, and Scharkow 2023; Hendrickx and Vázquez-Herrero 2024; Vázquez-Herrero, Negreira-Rey, and López-García 2022; Wirz et al. 2023). Examining videos published in 2022 by nine news outlets from Belgium, Spain, and the UK, Hendrickx and Vázquez-Herrero (2024) shed light on how news outlets continuously experiment with their content on TikTok. They revealed stronger tendencies towards additional visual elements and effects, such as transitions and stickers, exhibiting an on-screen journalistic presence for more current news reports on TikTok in contrast to Instagram (pp. 1062–1064, 1067). Wirz et al. (2023) compared the quality of content published by six major Swiss news outlets between February 2022 and January 2023 on TikTok and Instagram with that of their respective websites. The authors identified a complementary relationship between the content published on social media platforms and the news outlets' websites (Wirz et al. 2023, p.14). Although these studies offer valuable insights, a current and extensive characterization of news on TikTok published by German-speaking news outlets, i.e., in Germany, Austria, and Switzerland, is lacking.

Furthermore, whether and how news outlets continue to adhere to central journalistic notions, such as news values (Galtung and Ruge 1965), on TikTok remains unclear. While the concept of news values has been transferred to the digital realm (Harcup and O'Neill 2017; Kristensen and Bro 2024; Mast and Temmerman 2021), first applications to TikTok have focused on the journalists' perspective rather than on the content (Peterson-Salahuddin 2024). However, the content-level exploration of news (values) on TikTok is crucial to advance our knowledge on the production and distribution of news on social media platforms and to identify potential shifts in journalistic practice. Similarly, the analysis of such shifts also requires large-scale datasets and methods for automating the annotation of multimodal content – as of today, both are limited.

Data

The dataset "News on TikTok" consists of metadata and human-annotated data of the TikTok videos published by 18 major German-speaking news outlets in 2023. The initial scraping of the news outlets' TikTok channels for the video metadata took place in January 2024 and resulted in 8,623 videos. The human annotation of the videos with a codebook was subsequently conducted from September 2024 and February 2025. The overall video sample was divided into two subsamples, systematically stratified by news outlet. Two extensively trained annotators coded the first subsample between September and November 2024 (N_1 = 4,287), and the second subsample between December 2024 and February 2025 (N_2 = 4,336). The complete dataset (N = 8,623) was released open-access in April 2025 (Wedel et al. 2025).

Data Collection

The selection of major German-speaking news outlets on TikTok followed multiple steps with two exclusion criteria (see Appendix, Table 3). First, we generated a list of the top news outlets based on the percentage of weekly use and brand trust of news outlets in Germany, Austria, and Switzerland in 2023 (Newman et al. 2023, pp. 61, 77, 103). Second, we searched for their corresponding TikTok channels and excluded those not mainly producing news content (*content criterion*). Third, we scraped the metadata of the channels' videos published in 2023 with the TikTok Research API. The metadata includes:

- Unique video ID (M1)
- Unique ID of the used music (M2)
- Date and time that the video was created (M3)
- Like, view, comment, and share count at the time of scraping (M4-7)
- Hashtags used in the description (M8)

Fourth, we validated the data retrieved from the TikTok Research API by comparing the unique video IDs with those collected via a different scraping method. In this case, we used the Zeeschuimer browser extension (Peeters 2024), which collects the videos' links while scrolling through a channel page. As a result, we added 64 videos to the sample. In the same step, we excluded channels that had been insufficiently active throughout 2023 to avoid potential outliers. Based on the weekly activity distribution of the channels, we derived an *activity criterion* establishing a threshold to include only those channels that had posted at least two videos (90% of the outlets) in at least 26 weeks of 2023 (upper 70%). Overall, the scraping resulted in the metadata of 8,623 videos from 18 German-speaking news outlets in 2023 (see Table 1).

Data Annotation

In addition to the metadata, the dataset also contains human-annotated data on the videos. Two authors annotated the videos based on a codebook (see the dataset repository for the codebook). The codebook built on and adapted previously established operationalizations of news characteristics on TikTok and news values (Araujo and van der Meer 2020; Harcup and O'Neill 2017; Hendrickx and Vázquez-Herrero 2024; Schafraad and Zoonen 2020). In addition to four variables that served as an orientation for the annotators, i.e., the designated annotator (V1), video URL (V2), news outlet (V3), and country of origin (V4), the codebook encompassed:

- The *general classification* of the video as news (V5), i.e., whether the video dealt with news or other kinds of content (nominal variable with five levels; level 1 = news-related content). This variable was also used as a filter variable for videos containing, e.g., predominantly self-promotional content (level 2), branded content (level 3), or interactive community content (level 4). In such cases, the annotator was instructed to stop annotating the video, to leave the subsequent variables (V6-V18) blank and

News outlets by country	Samples		
	Complete Sample	Stratified Sample 1	Stratified Sample 2
Germany			
tagesschau	388	194	194
rtlaktuell	365	182	183
zeit	284	143	141
faz	277	138	139
Austria			
heute.at	1,090	543	547
derstandard	444	218	226
zeitimbild	424	210	214
kleinezeitung	389	194	195
kurier.at	365	181	184
krone.at	305	150	155
nachrichten.at	126	63	63
Switzerland			
20minuten	1,853	914	939
blick	754	376	378
watson_news	537	269	268
bluenews.ch	395	198	197
srfnews	349	174	175
neuezuercher-zeitung	157	79	78
tagesanzeiger	121	61	60
Total	8,623	4,287	4,336

Table 1: Distribution of the TikTok videos in the dataset by news outlet.

start with the next video. Similarly, videos that were unavailable under the provided link were also filtered out (level 5).

- *Formal variables* that covered the general visual set-up (V6; nominal variable with four levels), and the presence of visual, auditive, and interactive elements in the video (V7-9; nominal variables with two levels).
- *Content variables* to classify the video's news format type (V10; nominal variable with five levels), describe the video's general topic (V11; open text field), and note the presence of selected news values (V12-18; nominal variables ranging from two to four levels). The codebook included the news values geographical scope (geographical reference, national news, or foreign news), follow-up, positive news, negative news, conflict, scope (mention of directly or potentially involved people), and temporal scope (mention of whether the video reported on past, current/recent or future event).

Extensive annotation training took place and covered collectively working through the codebook, annotating, and discussing the annotations of a selected sample of videos. Datasheets were prepared and used for the annotation. After training, six pretests were conducted until the variables exhibited satisfactory intercoder reliability values of Krippendorff's $\alpha \geq 0.800$ (Krippendorff 2004, p. 429). For this purpose, we drew a random sample of TikTok videos from the 18 news outlets from January to March 2024. The number of videos subsequently annotated by two of the authors varied by pretest (ranging from 21-100 videos; M = 47.17; SD = 26.12; N = 286). Both of the annotating authors annotated each video in the pretest samples. Adaptations were iteratively made to the codebook between the pretests and included reformulations and cutting certain variables that either seldomly appeared in the pretests or did not reach satisfactory reliability. The final pretest encompassed 51 videos (see Appendix, Table 4). The intercoder reliability values were calculated using the rpackage *tidyComm* (Unkel, Haim, and Kobilke 2024).

The 8,623 videos were annotated in two phases (September to November 2024; December 2024 to February 2025). Each annotating author coded about half of the videos. To facilitate room for discussing potential questions and ensure consistent annotation quality, check-in meetings between the annotating authors and those who had developed the codebook were scheduled every two weeks, and an internal chat room was created for more ad hoc and shorter exchanges.

Data Availability

Adhering to the FAIR guidelines (Wilkinson et al. 2016), the dataset can be found in the open-access GESIS repository (Wedel et al. 2025): https://doi.org/10.7802/2863. The repository contains the dataset as an CSV file, the final codebook, and a data report that also refers to this dataset paper. The dataset is re-usable under the Creative Commons Attribution 4.0 International license.

Due to copyright protection, the dataset contains neither the TikTok videos nor the video descriptions. Alternatively, while acknowledging certain limitations of reproducibility in light of possible changes to the videos, such as removal or deletion, the dataset provides the unique video IDs and their URLs for traceability.

Preliminary Analysis

To provide a more detailed understanding of the dataset and to highlight its potential, we performed a preliminary descriptive analysis on the dataset (N = 8,623). The analyses were conducted with Python, using the packages *pandas* (The pandas development team 2024), *matplotlib* (Hunter 2007), and *WordCloud* (Mueller 2024). The following section provides an overview of the metadata (M3-8), the results of the general classification of the videos as news (V5), the general topics of the videos (V11), and a summary of the other formal and content variables (V6-10, V12-18).

Metadata

As Figure 1 shows, the news outlets in our dataset consistently published videos throughout 2023, with a slight upward trend over the year. Around July, the number of videos published per week peaked, while around the end of April and September, there were two dips. On average, a single news outlet published 9.21 videos per week (SD = 7.91); collectively, the news outlets averaged 165.83 videos per week (SD = 20.74).

The hashtags used most frequently by the news outlets point to three main tendencies (see Figure 2). First, the news outlets appear to strategically address TikTok's algorithmic selection. We infer this from their frequent use of *#fyp*, *#fy*, and *#fürdich* (English: *#foryou*), as abbreviations for the 'for

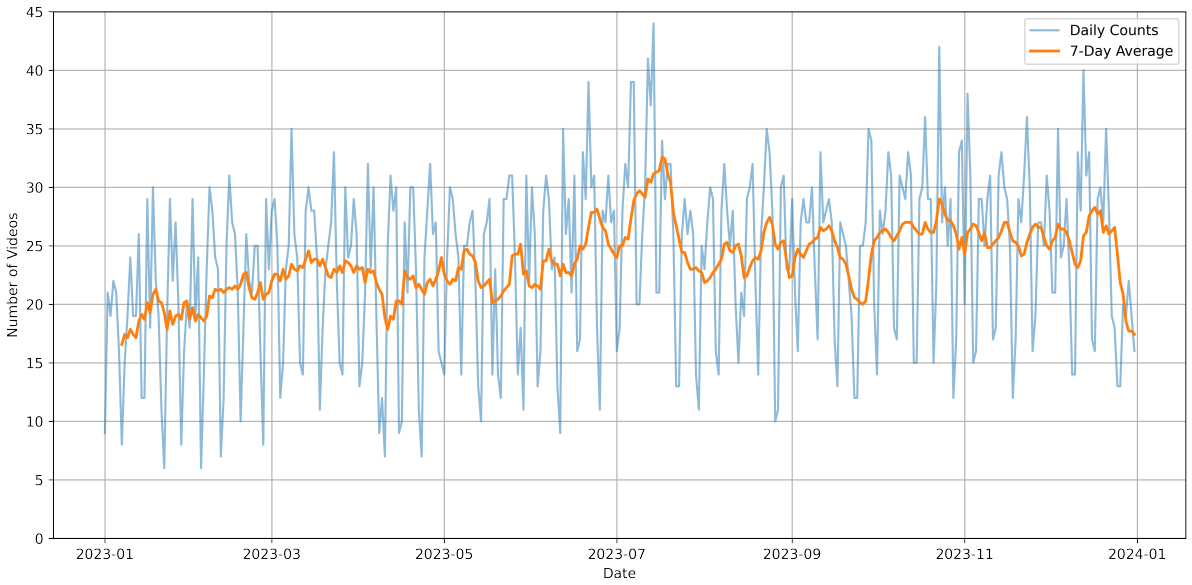

Figure 1: Distribution of the TikTok videos published by the news outlets across 2023 (N = 8,623).

you page' (the personalized feed subscribers see first when they open the app) and other hashtags referring to TikTok trends, such as *#lernenmittiktok* (English: *#learningwithtiktok*). These hashtags can be understood as "TikTok virality boosters" (Wedel, Hendrickx, and Mayer 2024, p. 23), potentially heightening the likelihood that the algorithm shares the videos in the feed on the 'for you page'. Second, the news outlets in our sample frequently signal the type of content they publish by using *#news* and *#nachrichten* (English: *#news*). Third, other frequent hashtags in the dataset refer to the names of news outlets, which could reflect efforts to improve (algorithmic) visibility and findability. The frequencies of these hashtags are, however, critically determined by the overall frequency of the news outlets' respective videos within the dataset. This results in, for example, the hashtags *#20min* from the Swiss news outlet *20minuten* (n = 1,853) or *#heutezeitung* and *#tageszeitungheute* from the Austrian news outlet *heute.at* (n = 1,090) to be more prominent in the word cloud.

The dataset also offers information on user engagement with the news outlets' TikTok videos (see Table 2). In 2023, on average (per video), the videos were viewed 48,878.32 times (SD = 191,810.48), received 4,523.64 likes (SD = 17,199.79) and 200.67 comments (SD = 1,655.49), and were shared 303.34 times (SD = 1,854.64). Notably, these numbers depend on the date of scraping (January 2024).

Figure 2: Word cloud of the hashtags used in the TikTok videos published by the news outlets in 2023 (N = 8,623).

	Mean	SD	Min	50%	Max
View	48,878.32	191,810.48	0	1,645	7,105,469
Like	4,523.64	17,199.79	1	707	821,805
Comment	200.67	1,655.49	0	22	137,075
Share	303.34	1,854.64	0	3	75,864

Table 2: Descriptive statistics of the TikTok videos' engagement metrics (N = 8,623; date of scraping: January 2024).

General Classification

The general classification of the videos (V5) looked at whether the videos contained news in a broader sense and determined whether the annotators should continue annotating the video or not (N = 8,623; see Appendix, Table 5). The preliminary analysis shows that most of the videos were classified as news-related (92.52%), which means that all formal and content variables (V6-V18) were annotated for 7,978 of the 8,623 TikTok videos. 2.17% were annotated as containing branded content, i.e., mainly promoting/advertising external events or products (n = 187), 1.99% as self-promotional content (n = 172), and 1.23% as mainly interactive community content (n = 106). In less than 3% of the cases, the video was excluded from further annotation because it was either unavailable (0.8%; n = 69) or its general classification was unclear (1.29%; n = 111).

General Topics

As part of the formal and content variables (V6-18) that were only annotated in the case that the video had been classified as news-related (V5; N = 7,978), the general topic of the video was briefly documented by the annotators (V11). The word cloud of the open responses in Figure 3 show that the videos dealt with a plethora of different topics. Skipping words describing the videos' formats, i.e., 'Beitrag' (English: post/article), 'Interview/s', 'Video', and 'Bericht' (English: report), we are able to identify different foci that can be investigated more in-depth with other and more sophisticated means of analysis, including:

- Locations, e.g., 'Schweiz' (English: Switzerland), 'Gaza', 'Wien' (English: Vienna), 'Türkei' (English: Turkey), and 'Österreich' (English: Austria).
- Populations, e.g., 'Mann' (English: man), 'Frau' (English: woman), 'Menschen' (English: people/humans), 'Schweizer' (English: Swiss), and 'Jugendliche' (English: adolescents/young people).
- Topics, e.g., 'Tipps' (English: tips), 'Studie' (English: study/report), 'Festival', 'Umfrage' (English: survey), and 'Polizei' (English: police).

Summary of Other Variables

The other formal and content variables (V6-10, V12-18) that were coded for 7,978 of the news-related videos can be summarized as follows (see Appendix, Table 5):

- Of all relevant videos in our sample, half of the visual set-ups were journalist-centric, i.e., a journalist or moderator was nearly constantly visible in the foreground (V6; 50.16%; n = 4,002).
- The most frequently used additional visual elements in the videos were text (V7b; 96.38%; n = 7,689), elements of the news outlet, such as a certain color schemes (V7e; 88.97%; n = 7,098), and images, including stickers, emojis, flags, and brand logos (V7a; 53.91%; n = 4,301).
- Audio-wise, the videos mainly contained a journalist's voice (V8a; 83.60%; n = 6,670) and music (V8d; 69.43%; n = 5,539).
- Interactive elements, i.e., the mention of user comments or the invitation to interact with the video or outlet on and outside of TikTok, were less present in the videos – with calls-to-interact on TikTok (V9b) used in 26.02% of the videos (n = 2,076).
- Regarding the type of news format (V10), most videos were classified as newsflashes or reports, as in informational videos that answered the main wh-questions (who, what, when, where) and potentially elaborated more on the topic (73.19%; n = 5,839). Other news formats in the dataset were interviews (11.11%; n = 886) and news summaries (10.08%; n = 804).
- The news values that were most frequently present in the videos were scope-related, including the temporal (V18; 86.21%; n = 6,878), geographic (V12; 85.75%; n = 6,841), and general scope (V17; 80.71%; n = 6,439).

Figure 3: Word cloud of the general topics annotated for the TikTok videos published by the news outlets in 2023 (N = 7,978).

Discussion

Summary

This paper introduces the dataset "News on TikTok" to explore pressing questions concerning the up-and-coming social media platform for the consumption of news and political information, TikTok, and the content that 18 major German-speaking news outlets produce for the platform. The preliminary analysis only scratches the surface of the dataset's breadth and depth. While beyond the remit of this paper, it could be easily augmented by comparing, for example, the different countries and news outlet types in the sample – but also beyond the sample. The dataset affords novel opportunities for researchers from various fields to explore empirical questions rooted in social science on the adaptation of traditional news outlets to TikTok's platform logic with respect to general characterizations of the news videos (visual, auditory, and interactive elements) and the presence of news values. Furthermore, human-annotated data of varying complexity provide a unique database to nurture and advance the contemporary discussion on automatized annotation techniques.

Limitations

There are limitations to the dataset and how it can be used. First, the metadata on user engagement depends on the time of scraping with the TikTok Research API (January 2024). This becomes more problematic the closer the publication date of a video is to the date of scraping. Furthermore, it should be noted that audits have generally identified stark fluctuations and instabilities in the engagement metrics provided by TikTok (Pearson et al. 2025). To account for these issues, we suggest that researchers interested in these metrics rescrape the metadata to receive stable estimates of video engagement.

Second, due to legal and ethical concerns, the dataset does not contain video files, descriptions, or transcripts. Instead, links to the videos are provided. Although this ensures the subject's right to deletion, the videos might not be available at a later point in time. However, with the respective videos being published from established news outlets, where each video presumably goes through multiple steps of quality control before upload, we do not expect a significant fraction of videos to become unavailable in the future, e.g., due to violations of TikTok's Terms of Service.

Finally, as a platform, TikTok has faced repeated threats of becoming unavailable in various markets (e.g., due to regulatory intervention). However, even in the unlikely case of widespread unavailability of the video data, the human annotations and the codebook would still maintain the relevance of the dataset. In any case, the dataset provides an extensive basis for understanding how news outlets produce news for an algorithmically-driven short video platform.

Conclusion

TikTok plays an increasingly relevant role in news consumption and distribution, but analyses on how traditional news outlets adapt their content to the social media platform and adhere to journalistic news values – as a central concept in journalism studies – are limited. Furthermore, large-scale datasets for the development of automated tools that can annotate the vast amount of multimodal content, such as on TikTok, are not available to date.

To address these research gaps and empirically explore the central questions surrounding algorithmically mediated news and political information on TikTok, we introduce "News on TikTok," an annotated dataset of the TikTok videos from 18 major German-speaking news outlets published in 2023 (N = 8,623). Augmenting the videos' metadata with the human-annotated data of 25 variables, including formal and content categories derived from journalism studies and news values, the contributions of the dataset can be summarized as follows: First, the dataset provides the basis for a more extensive systematic analysis and comparison of TikTok news videos. It builds on and applies operationalizations established in previous studies that have explored characteristics of news videos on TikTok, which enables comparability. Second, it provides a large and validated amount of human-annotated data that can be used to develop and validate automated tools for analyzing visual and auditory video elements. Given the increasing amount of (news) content uploaded to platforms, the development of such automated approaches is increasingly relevant in social science research. Third, the dataset paper offers a comprehensive guide to generate datasets to explore similar research interests concerning news on TikTok or any video-based platform with different geographic contexts, such as the US, or thematic foci.

Acknowledgements

We would like to acknowledge funding by the Federal Ministry of Education and Research of Germany (BMBF) under grant no. 16DII131 (Weizenbaum-Institut für die vernetzte Gesellschaft – Das Deutsche Internet-Institut). Furthermore, we would like to acknowledge the internal funding program of the Weizenbaum-Institut for interdisciplinary short projects 2024-25, which supported this research within the project "Labeling of TikTok data for social science research: Comparing large language models to expert human annotators."

References

Araujo, T.; and van der Meer, T. G. 2020. News values on social media: Exploring what drives peaks in user activity about organizations on Twitter. *Journalism*, 21(5): 633–651.

Deutsche Forschungsgemeinschaft. 2022. Guidelines for Safeguarding Good Research Practice. Code of Conduct. Report, Deutsche Forschungsgemeinschaft.

Galtung, J.; and Ruge, M. H. 1965. The Structure of Foreign News: The Presentation of the Congo, Cuba and Cyprus Crises in Four Norwegian Newspapers. *Journal of Peace Research*, 2(1): 64–90.

Harcup, T.; and O'Neill, D. 2017. What is News?: News values revisited (again). *Journalism Studies*, 18(12): 1470–1488.

Hase, V.; Boczek, K.; and Scharkow, M. 2023. Adapting to Affordances and Audiences? A Cross-Platform, Multi-Modal Analysis of the Platformization of News on Facebook, Instagram, TikTok, and Twitter. *Digital Journalism*, 11(8): 1499–1520.

Hendrickx, J.; and Vázquez-Herrero, J. 2024. Dissecting Social Media Journalism: A Comparative Study Across Platforms, Outlets and Countries. *Journalism Studies*, 25(9): 1053–1075.

Hunter, J. D. 2007. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, 9(3): 90–95.

Krippendorff, K. 2004. Reliability in Content Analysis: Some Common Misconceptions and Recommendations. *Human Communication Research*, 30(3): 411–433.

Kristensen, L. M.; and Bro, P. 2024. News values in a digital age- Intra-media, inter-media, and extra-media platforms. *Journalism*, 25(4): 819–836.

Leppert, R.; and Matsa, K. E. 2024. More Americans – especially young adults – are regularly getting news on TikTok. Technical report, Pew Research Center.

Mast, J.; and Temmerman, M. 2021. What's (The) News? Reassessing "News Values" as a Concept and Methodology in the Digital Age. *Journalism Studies*, 22(6): 689–701.

Mueller, A. C. 2024. Wordcloud.

Newman, N.; Fletcher, R.; Eddy, K.; Robinson, C. T.; and Nielsen, R. K. 2023. Reuters Institute Digital News Report 2023. Technical report, Reuters Institute for the Study of Journalism.

Newman, N.; Fletcher, R.; Robertson, C. T.; Arguedas, A. R.; and Nielsen, R. K. 2024. Reuters Institute Digital News Report 2024. Technical report, Reuters Institute for the Study of Journalism.

Pearson, G. D. H.; Silver, N. A.; Robinson, J. Y.; Azadi, M.; Schillo, B. A.; and Kreslake, J. M. 2025. Beyond the margin of error: A systematic and replicable audit of the TikTok research API. *Information, Communication & Society*, 28(3): 452–470.

Peeters, S. 2024. Zeeschuimer.

Peterson-Salahuddin, C. 2024. News for (Me and) You: Exploring the Reporting Practices of Citizen Journalists on TikTok. *Journalism Studies*, 25(9): 1076–1094.

Schafraad, P.; and Zoonen, W. V. 2020. Reconsidering churnalism: How news factors in corporate press releases influence how journalists treat these press releases after initial selection. *Communications*, 45(s1): 718–743.

The pandas development team. 2024. pandas-dev/pandas: Pandas (v2.2.3). 10.5281/zenodo.13819579. Accessed: 2025-03-01.

Unkel, J.; Haim, M.; and Kobilke, L. 2024. tidycomm: Data Modification and Analysis for Communication Research.

Vázquez-Herrero, J.; Negreira-Rey, M.-C.; and López-García, X. 2022. Let's dance the news! How the news media are adapting to the logic of TikTok. *Journalism*, 23(8): 1717–1735.

Wedel, L.; Hendrickx, J.; and Mayer, A.-T. 2024. What occurs as #news on TikTok? A Computational Approach. In *Companion Publication of the 16th ACM Web Science Conference*, 22–23.

Wedel, L.; Mayer, A.-T.; Batzner, J.; and Hendrickx, J. 2025. News on TikTok: An Annotated Dataset of TikTok Videos from German-Speaking News Outlets in 2023. *GESIS Data Archive*.

Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L. B.; Bourne, P. E.; Bouwman, J.; Brookes, A. J.; Clark, T.; Crosas, M.; Dillo, I.; Dumon, O.; Edmunds, S.; Evelo, C. T.; Finkers, R.; and Mons, B. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1): 160018.

Wirz, D. S.; Zai, F.; Vogler, D.; Urman, A.; and Eisenegger, M. 2023. Die Qualität von Schweizer Medien auf Instagram und TikTok: Jahrbuch Qualität der Medien Studie 2 / 2023. Technical report, fög - Forschungszentrum Öffentlichkeit und Gesellschaft/University of Zurich.

Ethics Checklist

1. For most authors...

   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, this dataset encompasses metadata and annotated data of videos from German-speaking news outlets that were published on TikTok in 2023. The dataset does not contain any personal data. For more details, see the Data and Ethical Statement sections.

   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes, see the Abstract and Introduction sections.

   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, see the Data Collection and Data Annotation sections.

   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? NA

   (e) Did you describe the limitations of your work? Yes, see the Limitations section.

   (f) Did you discuss any potential negative societal impacts of your work? Yes, see the Ethical Statement section.

   (g) Did you discuss any potential misuse of your work? Yes, see the Ethical Statement section.

   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, see the Ethical Statement section.

   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes, we have read the ethics review guidelines and ensured that our

paper conforms to them. For more details, see the Ethical Statement section.

2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? NA
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? NA
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? NA
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? NA
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes, see the Data Collection section.
   (b) Did you mention the license of the assets? NA
   (c) Did you include any new assets in the supplemental material or as a URL? Yes, see the URL to the open-access repository for the codebook.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? NA
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? NA
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? Yes, see the Dataset Availability section.
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (Gebru et al., 2021)? No, we did not create an additional datasheet because this paper covers the aspects recommended in the Datasheet for Datasets (Gebru et al., 2021).
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? Yes, we have included the codebook in the Supplementary Material. For more details, see the Data Annotation section.
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? Yes, see the Data Annotation and Ethical Statement sections.
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA
   (d) Did you discuss how data is stored, shared, and deidentified? Yes, see the Dataset Availability section.

Ethical Statement

The dataset is based on publicly available data, i.e., TikTok videos published by major German-speaking news outlets in 2023. To comply with copyright laws, we only collected and provide the unique video IDs and generated video URLs in the dataset. We did not seek formal approval from the Institutional Review Board (IRB) at our institutions, but oriented our research according to the basic guidelines for safeguarding good research practice in Germany by the German Research Foundation (Deutsche Forschungsgemeinschaft 2022), for the following reasons: We look at publicly available data that has gone through professional journalistic editing and the upload content moderation of the TikTok platform. We also comply with copyright laws regarding the videos, and do not save personally identifiable information on the level of single individuals in our dataset. In this sense, we perceive a low danger of potential negative societal impacts from our dataset but instead see the annotated data as a means to investigate and reflect on the distribution of news on TikTok by established news outlets, which in itself could drive a positive societal impact.

Appendix

News outlet	TikTok channel	Active weeks	Exclusion
Germany			
Tagesschau	tagesschau	52	
zdf heute	zdfheute.nachrichten	0	Activity
RTL aktuell	rtlaktuell	52	
Spiegel TV	spiegeltv	5	Activity
Zeit	zeit	49	
n-tv	ntv.de	21	Activity
Stern	stern_de	7	Activity
t-online	tonline.de	0	Activity
jetzt.de (SZ)	jetzt.de	14	Activity
FAZ	faz	52	
Austria			
ORF - Zeit im Bild	zeitimbild	52	
ORF - radiowienheute	orfradiowienheute	51	Content
Kronen Zeitung	krone.at	52	
ServusTV On	servustvon	52	Content
Heute	heute.at	52	
oe24	oe24at	16	Activity
Kleine Zeitung	kleinezeitung	52	
Der Standard	derstandard	51	
Kurier	kurier.at	52	
MeinBezirk	bezirksrundschau	16	Activity
Die Presse	diepressecom	9	Activity
OÖ Nachrichten	nachrichten.at	31	
Salzburger Nach.	salzburgernachrichten	10	Activity
Switzerland			
20 Minuten	20minuten	52	
Blick	blick	52	
Tagesanzeiger	tagesanzeiger	36	
SRF News	srfnews	52	
Blue News	bluenews.ch	48	
nau.ch	nau.ch	0	Activity
NZZ	neuezuericherzeitung	45	
Watson	watson_news	52	

Table 3: Overview of the news outlet selection process for the dataset.

Variable label	Measurement level	N of levels**	Krippendorff's alpha
Metadata			
M1_video_id*	nominal	-	-
M2_music_id*	nominal	-	-
M3_create_time*	date	-	-
M4_like_count*	metric	-	-
M5_view_count*	metric	-	-
M6_comment_count*	metric	-	-
M7_share_count*	metric	-	-
M8_hashtags*	nominal	-	-
Codebook			
V1_Coder_ID*	nominal	2	-
V2_video_link*	nominal	-	-
V3_news_outlet*	nominal	18	-
V4_country_of_origin*	nominal	3	-
V5_news_classification	nominal	5	0.793
V6_visual_set-up	nominal	4	0.893
V7a_images	nominal	2	0.913
V7b_text	nominal	2	1.000
V7c_screenshots_and_recordings	nominal	2	0.935
V7d_graphs_maps_infoanimations	nominal	2	1.000
V7e_news_brand_elements	nominal	2	1.000
V7f_outlet-related_endcard	nominal	2	0.956
V8a_journalists_moderators_voice	nominal	2	1.000
V8b_other_peoples_voice	nominal	2	0.949
V8c_voice_other_effect	nominal	2	0.874
V8d_music	nominal	2	0.953
V8e_ambient	nominal	2	0.935
V9a_mentions	nominal	2	0.000
V9b_call_to_interact_on_tiktok	nominal	2	0.913
V9c_call_to_interact_outside_of_tiktok	nominal	2	0.816
V10_news_format_type	nominal	5	1.000
V11_general_topic*	open text field	-	-
V12_geographic_scope	nominal	3	0.966
V13_follow_up	nominal	2	1.000
V14_positive_news	nominal	2	1.000
V15_negative_news	nominal	2	0.911
V16_conflict	nominal	2	0.841
V17_scope	nominal	2	1.000
V18_temporal_scope	nominal	4	0.824

* These variables were omitted from the intercoder reliability test. They are either metadata, were added into the codebook and datasheets for the annotator's orientation, or required open text field responses.

** The code '-99' for unclear cases was excluded from the count of levels.

Table 4: Overview of the dataset's variables (metadata and codebook) with the intercoder reliability results from the final pretest ($N = 51$).

Variable label	Levels with n (%)						
	0	1	2	3	4	77	-99
V5_news_classification	-	7,978 (92.52)	172 (1.99)	187 (2.17)	106 (1.23)	69 (0.8)	111 (1.29)
V6_visual_set-up	-	4,002 (50.16)	1,577 (19.77)	2,398 (30.06)	-	-	1 (0.01)
V7a_images	3,677 (46.09)	4,301 (53.91)	-	-	-	-	-
V7b_text	289 (3.62)	7,689 (96.38)	-	-	-	-	-
V7c_screenshots	7,180 (90.0)	798 (10.0)	-	-	-	-	-
V7d_graphs_maps	7,410 (92.89)	567 (7.11)	-	-	-	-	-
V7e_news_brand	880 (11.03)	7,098 (88.97)	-	-	-	-	-
V7f_endcard	6,257 (78.43)	1,721 (21.57)	-	-	-	-	-
V8a_moderator_voice	1,308 (16.4)	6,670 (83.6)	-	-	-	-	-
V8b_other_voice	6,022 (75.48)	1,956 (24.52)	-	-	-	-	-
V8c_voice_effect	5,287 (66.27)	2,691 (33.73)	-	-	-	-	-
V8d_music	2,439 (30.57)	5,539 (69.43)	-	-	-	-	-
V8e_ambient	5,502 (68.96)	2,476 (31.04)	-	-	-	-	-
V9a_mentions	7,811 (97.91)	167 (2.09)	-	-	-	-	-
V9b_tiktok_interact	5,902 (73.98)	2,076 (26.02)	-	-	-	-	-
V9c_external_interact	6,603 (82.77)	1,375 (17.23)	-	-	-	-	-
V10_news_format	-	260 (3.26)	5,839 (73.19)	804 (10.08)	886 (11.11)	181 (2.27)	8 (0.1)
V12_geographic	1,133 (14.2)	4,453 (55.82)	2,388 (29.93)	-	-	-	4 (0.05)
V13_follow_up	7,952 (99.67)	26 (0.33)	-	-	-	-	-
V14_positive_news	7,543 (94.55)	435 (5.45)	-	-	-	-	-
V15_negative_news	6,219 (77.95)	1,759 (22.05)	-	-	-	-	-
V16_conflict	6,856 (85.94)	1,122 (14.06)	-	-	-	-	-
V17_scope	1,539 (19.29)	6,439 (80.71)	-	-	-	-	-
V18_temporal	1,097 (13.75)	158 (1.98)	6,449 (80.83)	271 (3.4)	-	-	3 (0.04)

Table 5: Count and frequency of the dataset's annotated variables (V5 with N = 8,623; V6-10, V12-18 with N = 7,978).