Decoding the Algorithm:
Evaluating TikTok's Influence as a News Source

Johanna Lee
Jyontika Kapoor
Tayae Rogers
Audrey Ming Wai Yip
Jenni Yu

CS315-02, Group 3
29 March 2024

Contents

1 Introduction

In the dynamic landscape of social media, TikTok has emerged as a powerful force, captivating millions with its short-form video content. As its popularity surges, researchers are increasingly turning their attention to the prevalence of news on this platform. According to the Pew Research Center, approximately one-third of adults under the age of thirty in the United States rely on TikTok for news consumption [9], underscoring its importance as a possible news source for younger demographics especially. However, despite its influence, TikTok's recommendation system remains opaque, often described as a "black box," which poses a challenge in understanding how news content is surfaced and disseminated on the platform [3]. Unlike other social media platforms, TikTok's algorithm-driven recommendation system operates behind the scenes, crafting personalized content tailored to individual user preferences and interactions. Consequently, the extent to which news creators are promoted by the algorithm is uncertain.

This paper aims to either corroborate or refute the finding of a preceding study which discovered that among US-based demographics, there is a significant lack of news dissemination on TikTok [7]. The methodological questions that guide our research are as follows: what methods are effective at identifying "news" content on TikTok? How do these methods contribute to our understanding of the amount and types of news topics on the platform? We hypothesize that clustering (H1) will be useful for understanding different types of news topics, and cosine similarity (H2) will assist us in quantifying the amount of news topics.

To address these questions, we leverage de-identified user data obtained from TikTok to measure the similarity between viewed videos and news-related hashtags, news accounts, and New York Times content. We isolate the period from October 7, 2023 to December 7, 2023 to investigate whether the world events concerning Israel-Palestine had an impact on the volume of news content circulated on TikTok. Overall, this research seeks to shine light into the platform's evolving role in the contemporary media landscape and inform discussions surrounding its impact on information dissemination.

2 Literature Review

TikTok's personalized content delivery is a distinctive feature of the platform. However, the mechanics behind TikTok's video recommendations remain largely unknown, prompting interest and inquiry among researchers [3]. Understanding how TikTok recommends videos, whether through user interactions such as likes, follows, or other algorithmic mechanisms, is an ongoing area of investigation attracting increasing attention from the computer science community [8].

While many researchers have explored the phenomenon of personalization on TikTok, our focus is specifically on its implications for news dissemination. Some sources say news consumption via social media platforms is becoming increasingly prevalent among younger demographics, ensuring they stay informed about current events that might otherwise be overlooked [2]. However, other researchers argue that not all users on these platforms are equally exposed to news content [12]. This lack of exposure may be because some users intentionally avoid news due to news fatigue or general disinterest [11]. Additionally, dynamic recommendation systems like TikTok prioritize content based on user content and interactions, possibly filtering our news sources in favor of other types of content, exacerbating the lack of news for some users [5].

However, recommendation systems are susceptible to popularity bias, so they may promote incidental exposure among users who do not show interest in news content [1], particularly during times of global events when popular or viral news content may circulate. Multiple studies show that TikTok has been used to mobilize international solidarity and disseminate information during Israel-Palestine world events [4, 13]. Therefore, questions about news content on TikTok are complex, emphasizing our interest in examining the extent to which its recommendation system exposes users to news content—and whether this exposure coincides with ongoing global events.

3 Data and Methods

3.1 Data Collection

Metadata was obtained from three users, each acquiring their data directly from TikTok. From the metadata, video browsing history and a following list was extracted for each user. The video browsing history comprised dates and corresponding links to viewed videos, lacking substantial

information about the videos themselves. Subsequently, we utilized the Pyktok Python module [6] to augment the information for each video within the period from October 7, 2023, to December 7, 2023. Using the default features of Pyktok, data was collected on video_id, video_timestamp, video_locationcreated, and video_description. Further details on additional features can be found in Appendix A1.

Additionally, modifications were made to the code to extract suggested_words, a list of keywords describing video content. This variable proved beneficial as only 31 percent of the collected videos had available descriptions. In total, data was gathered from 163,753 videos, with 154,256 from user one, 9,014 from user two, and 483 from user three. Notethat certain videos, no longer available on TikTok at the time of data collection, were identified during the Pyktok scraping process and subsequently excluded from the final datasets along with their respective video counts.

After the data collection process, the data from the three users were aggregated. Subsequently, the aggregated dataset was filtered to identify videos possibly related to news. These videos were identified using three distinct methods: (i) comparison with an aggregated list of news-related hashtags, (ii) comparison with news-related accounts, and (iii) comparison with articles from The New York Times (NYT).

3.2 Filtering by Comparison with News-Related Hashtags

For the identification of news-related content through hashtags, the hashtags associated with each video were compared to a list of 147 hashtags compiled by a Wellesley College 300-level computer science class. If a video's description contained one or more hashtags from the class list, it was categorized as news-related. Through this process, 86 TikTok videos were identified as potentially containing news content.

3.3 Filtering by Comparison with News-Related Accounts

For identifying news-related videos based on the poster, a comparison was conducted between the poster of each video and a list of 490 news accounts compiled by the same 300-level computer

science class at Wellesley College. If a video was posted by an account from the class list, it was categorized as news-related. Through this method, 342 TikTok videos were identified as potentially constituting news content.

3.4 Filtering by Similarity with New York Times Headlines

The third and final method involves comparing the videos with New York Times headlines. For this, select components of Pyktok metadata (like country, suggested words, and description) along with cosine similarity were used to filter the videos, following the process outlined in Figure 1 below.

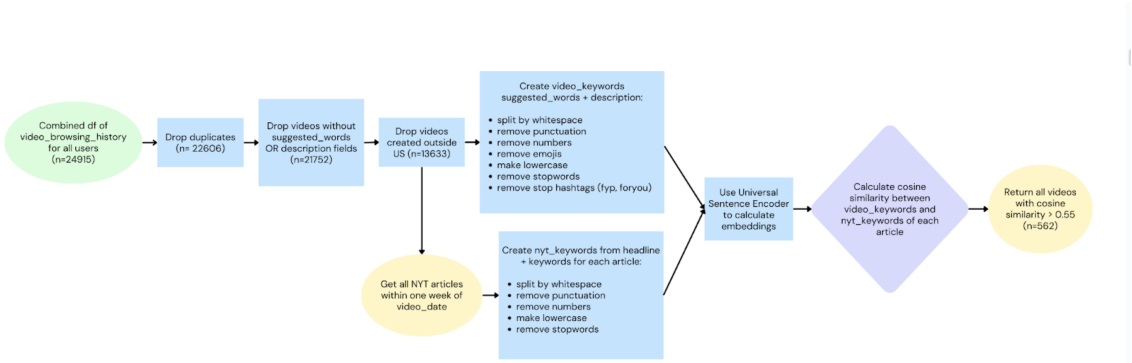

Figure 1: The flowchart illustrates the process of filtering videos by comparing them to NYT articles. This included excluding non-US videos and those without suggested words or descriptions, generating keyword lists from both videos and NYT articles, and calculating cosine similarity using Universal Sentence Encoder embeddings. Comparisons were made between video keywords and articles published within a week of the video's upload, with videos having cosine similarity greater than or equal to 0.55 considered for transcription.

Then, for each video, list of keywords was compiled by merging the suggested words and description, while omitting common hashtags like 'fyp' and 'for you,' which are incompatible with NYT articles. Furthermore, given the disparities between the textual content on TikTok (such as descriptions) and NYT articles, we also generated a list of keywords for each NYT article, extracted from the headline and abstract. The use of these keywords aimed to enhance the effectiveness of using cosine similarity as a metric, enabling more accurate comparisons between TikTok videos and NYT articles.

Lastly, we utilized the Universal Sentence Encoder to compute embeddings for each list of keywords. Subsequently, for each video, cosine similarity was calculated between its keyword list and the keyword list of each NYT article published within one week of the video's upload date (three days before, the upload day, and three days after). This approach was chosen as it yielded a higher median cosine similarity (0.315) compared to solely comparing with articles on the video upload date (0.249). As a result, 516 videos with a cosine similarity of 0.55 or higher were identified for transcription.

3.5 Transcription

From our three filtering processes (by hashtags, by posting account, and by similarity with NYT headlines), 834 unique videos were identified as possibly related to news. Among these, 450 videos were randomly chosen for transcription. Note that we opted for sampling due to computational constraints.

To transcribe the selected videos, the .mp4 files were downloaded through Pyktok [6]. Then, each .mp4 file was passed to librosa, an audio processing package, to load and resample the audio at 16000 Hz. Afterward, Whisper, an open-source neural network specialized in automatic speech recognition [10], was utilized. Before the audio was passed to Whisper, it underwent processing by the Whisper processor, which includes both the feature extractor (responsible for pre-processing raw audio) and the tokenizer (used to decode transcriptions into words). Finally, the pre-processed audio was fed into the WhisperForConditionalGeneration model, which generated and returned IDs corresponding to tokens, resulting in the production of a .txt file containing the transcription for each video.

3.6 Comparison of Transcriptions to NYT articles

From the transcribed videos, transcripts in languages other than English were excluded. Additionally, transcripts containing fewer than 18 words (the maximum NYT headline length) were filtered out. This decision was driven by the sensitivity of cosine similarity to length, as it is unlikely that transcripts with only a few words would contain substantial meaningful news content.

Then, the remaining 227 transcriptions were compared with NYT headlines, abstracts, and lead paragraphs to assess the effectiveness of the filtering methods. Similar to the filtering step, this comparison used cosine similarity of Universal Sentence Encoder embeddings, with no pre-processing applied to either the video transcription or article headline.

Additionally, considering the computational expense of comparing full transcripts to entire headlines, abstracts, and lead paragraphs, we sought to determine if this approach would outperform comparing keyword fields. Hence, we constructed keyword fields for the NYT articles, as detailed in Section 3.4. Similarly, we created transcription keywords by eliminating stopwords and punctuation. Subsequently, we calculated the cosine similarity between the Universal Sentence Encoders of the cleaned transcription and NYT keyword and headline data.

3.7 Qualitative Analysis

To validate the findings regarding news-related videos identified through cosine similarity, a qualitative examination was conducted on those with the highest cosine scores. The definition of news for this analysis encompassed videos reporting or commenting on current events. Each video underwent manual analysis by three individuals, who categorized them as 'Yes,' 'No,' or 'Maybe' based on the aforementioned definition. In cases of conflicting categorizations, discussions were held to reach a consensus. Additionally, a subset of news videos pertaining to the events in Israel-Palestine were identified.

4 Results

Recall that our primary research questions revolved around the identification of news on TikTok, the types of news showcased on the platform from October 7, 2023, to December 7, 2023, and the extent of its presence. Out of our original dataset comprising 24,915 videos, we pinpointed 834 unique videos as news-related through our three filtering methods. Subsequently, we analyzed this dataset using three approaches: (i) cosine similarity between transcripts and The New York Times, (ii) clustering, and (iii) qualitative analysis of transcripts.

4.1 Cosine Similarity Between Transcripts

Figure 2 below presents the distribution of cosine similarities computed for different groups. Across the four comparisons of the TikTok transcripts of videos identified as news-related, all yielded right-skewed cosine similarity distributions with medians ranging between 0.30 and 0.35. Notably, the median cosine similarity for comparing video keywords to NYT keywords across all videos was very similar, at 0.305. Consequently, we conclude that using cosine similarity to compare transcripts with the NYT does not appear to be any more effective in identifying news than using cosine similarity to compare video keywords to NYT keywords.

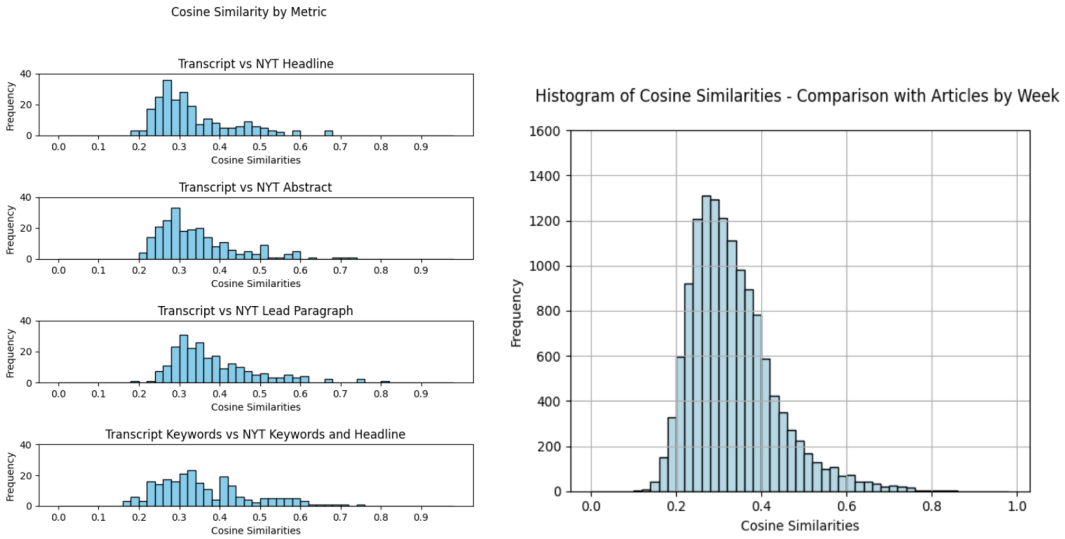

Figure 2: $Left$: Histograms of the cosine similarities based on the video transcriptions, for videos already identified as potentially news-related. From top to bottom, their medians are: 0.304, 0.319, 0.353, and 0.332 $Right$: Histogram of the cosine similarities between video keywords and NYT keywords, for all videos. Its median is 0.315.

4.2 Clustering

Figure 3 depicts the results of K-Means clustering, applied to the 516 TikTok videos initially identified as news through similarity with NYT headlines. Among the eight clusters, seven exhibit cohesive themes as outlined in Appendix A2. Notably, six out of these seven clusters feature content related to cultural topics or soft news, such as 'The Hunger Games' or 'Thanksgiving.' The sole cluster related to world news is cluster 2 on Israel-Palestine, which comprises 71 videos, constituting 13.76 percent of the dataset.

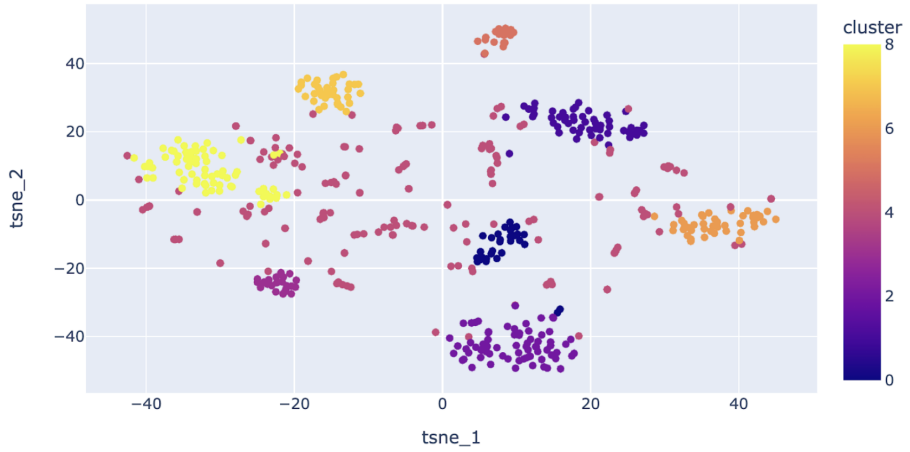

Figure 3: K-means Clustering of Embeddings of TikTok Videos Filtered by NYT Headline Comparison (k=8). Seven clusters are distinct and cohesive (e.g. the orange cluster 7 is about K-pop and the yellow cluster 8 is about Taylor Swift), while one cluster (the pink cluster 4) lacks a cohesive theme.

4.3 Qualitative Analysis

Recall that for the qualitative examination of the transcripts, non-English content and transcripts less than 18 words long were filtered out. To human-verify whether the videos were news-related, we categorized them into three groups: 'Yes,' 'Maybe,' and 'No.' Through this process, only 72 out of the 227 videos were identified as definitively or possibly news-related, amounting to 31.7 percent. As a result, we observed that even during our selected time period, which was chosen because we hypothesized finding a substantial amount of news content, the actual news content constituted a

very small percentage of our users' TikTok feeds.

Notably, in the human-verified review of the transcripts, it was found that 37 videos (accounting for 51.34 percent of the sampled news-related transcripts) were related to Israel-Palestine. Although the extent of news on TikTok during our selected time period cannot be compared to another time period within the scope of this study, the current analysis suggests that a majority of the news content users received pertained to Israel-Palestine.

5 Discussion

While a considerable amount of the news content identified was related to Israel-Palestine, overall, news constituted a small portion of TikTok content during the period between October 7, 2023, and December 7, 2023. These findings could indicate that there is indeed a limited amount of news content on TikTok, which aligns with the findings of the original study [7]. However, this contrasts with a report indicating that a significant portion of TikTok users – particularly young people – claim to obtain news from the platform [9]. This suggests a discrepancy between how researchers and young people define and measure news.

Through qualitative research, the difficulty of defining news became apparent. While manually tagging transcripts as news, numerous videos posed challenges in categorization. A wide-reaching definition of news was employed, yet many transcripts could have been alternatively tagged as current events-related, on-the-ground coverage, or commentary. To elaborate on the examination framework, 20 transcripts were designated as 'Maybe' news. This category encompassed mixed content, such as commentary on non-news events with a reference to world events like Israel-Palestine or on politicians. These transcripts were included in the pool due to our selection of hashtags. Additionally, among the non-news content, we observed an intriguing presence of Taylor Swift lyrics. These lyrics were initially identified as potentially news through the keyword comparison to NYT keywords, specifically resembling style and opinion articles.

In some cases, decisions on categorizing transcripts as news hinged on the political context at the time and the speaker's identity – information we lacked. For instance, a transcript discussing being

pro-life raised uncertainty about its news value, as political opinions alone don't guarantee newsworthiness. Further context, like whether it was in response to an event, was deemed necessary. However, upon viewing the video, it became clear that the speaker was Nikki Haley, a Presidential candidate, thus qualifying it as news. This underscores the limitation that categorizing videos based solely on transcripts is not always feasible.

Note that measurements of news – based on Hagar's and Diakopoulus' [7] – may have been flawed. From personal experiences, it's observed that a significant portion of news-related TikTok content originates from non-traditional news sources and is not presented in a traditional reporting style. Reflecting the egalitarian nature of TikTok, news often consists of individual users discussing current events and expressing their opinions. It's plausible that this is the type of news young people refer to when claiming to get news from TikTok. Furthermore, this form of news may not align well with metrics, as it often originates from non-traditional sources and may not be structured similarly to content from traditional news outlets, particularly in terms of language and format, as measured by cosine similarity.

In addition to the findings, the current analysis also presents further research opportunities. Future research can extend beyond textual analysis and incorporate the visual component of the videos. Tools such as graph machine-learning techniques can uncover hidden patterns within TikTok's content ecosystem. Another promising direction is to expand upon the clustering analysis: by examining how New York Times keywords fit into existing clusters and investigating how clusters evolve over different time frames, researchers can gain insights into the dynamic nature of news-related content on TikTok. Additionally, exploring variations in clusters across different user demographics could offer valuable insights into platform usage patterns.

To address the concern of having limited data within a constrained time frame, future work can expand the current analysis by comparing our selected time period to another. This comparison can focus on different time frames excluding incidents like Israel-Palestine to eliminate the impact on news types introduced by major world events. It can be used to investigate the association between news type distribution and significant events, providing insights into whether and how Israel-Palestine influenced the amount of total news content on TikTok.

Finally, redefining the concept of news within the context of TikTok can also present an innovative research opportunity. As our study revealed, traditional definitions may not fully capture the diverse forms of information dissemination and user-generated content on the platform. Future investigations should consider broader conceptual frameworks that encompass the range of content types and potentially tailor the definition of news to align with individual user preferences and backgrounds.

6 Conclusion

In our investigation into TikTok's function as a news source, it was found that despite the platform's popularity among younger demographics, news content occupies only a small fraction of users' feeds during our selected timeframe. Our analysis suggests that TikTok's algorithm assigns low priority to news dissemination, despite the global significance of events such as events relating to Israel-Palestine conflict.

In conclusion, while this study sheds light on TikTok's role as a news source, there remains a rich area for exploration. By addressing the limitations of the current approach while embracing innovative methodologies and redefining key concepts, researchers can continue to unravel the complexities of news consumption in this growing era of TikTok.

Appendix A1

Features collected through Pyktok. Bold features were used.

- **video_id**
- **video_timestamp**
- video_duration
- **video_locationcreated**
- **suggested_words**
- video_likecount
- video_sharecount
- video_commentcount
- video_playcount
- **video_description**
- **video_is_ad**
- video_stickers
- author_username
- author_name
- author_followercount
- author_followingcount
- author_heartcount
- author_videocount
- author_diggcount
- author_verified

Appendix A2

Table 1: Cluster Information

Cluster Number, (Color), Location	**Theme**	**Video Keywords of Sample in Cluster**
0 (blue) middle	New York	['much', 'rent', 'door', 'wont', 'close', 'nyc', 'rent', 'apartment', 'viral', 'hurricane', 'hurricanebridgit-mendler']
1 (bright purple) top right	Halloween	['happy', 'halloween', 'guyssss', 'homelessant', 'homelessantmeme', 'halloween', 'halloween', 'antmeme']
2 (purple) bottom middle	Israel-Palestine	['day', 'ceasefire', 'yet', 'theyre', 'still', 'bmbing', 'gaza', 'youre', 'pro', 'isnotreal', 'youre', 'actually', 'sick', 'freepalestine']
3 (dark pink) bottom left	The Hunger Games	['tigris', 'hunger', 'games', 'lucy', 'gray', 'snow', 'hunger', 'games', 'hunger', 'games', 'songbirds', 'snakes', 'hunger', 'games', 'happened', 'lucy', 'gray', 'hunger', 'games', 'ballad', 'songbirds', 'ballad', 'songbirds', 'snakes', 'hunger', 'games', 'explained', 'president', 'snow', 'put', 'together', 'hungergames', 'balladofsongbirdsandsnakes', 'presidentsnow', 'lucygraybaird', 'endingexplained', 'movieanalysis', 'booktok']
4 (pink) scattered middle	NA	['jimmy', 'carter', 'young', 'jimmy', 'rosalynn', 'carter', 'hospice', 'jimmy', 'carter', 'hospice', 'jimmycarter', 'jimmy', 'rosalynn', 'carter', 'young', 'jimmy', 'carter', 'rosalynn', 'cnn', 'jimmy', 'carter', 'jimmy', 'carter', 'presidency', 'jimmy', 'carter', 'palestine', 'greenscreenvideo', 'love', 'jimmy', 'carter', 'roslynn', 'carter', 'potus']
5 (grapefruit) top middle	TV shows	['informant_part', 'antfarm', 'disney']
6 (orange) far right	Thanksgiving	['best', 'mexican', 'thanksgiving', 'turkey', 'thanksgiving', 'mexican', 'turkey']
7 (light orange) top left	K-pop	['serious', 'would', 'think', 'finna', 'actually', 'debut', 'kpop', 'jyp', 'sm', 'yg', 'hybe', 'cube', 'blackpink', 'bts', 'twice', 'straykids', 'gidle']
8 (yellow)	Taylor Swift	['taylor', 'swift', 'cats', 'taylor', 'swift', 'productions', 'taylor', 'swift', 'productions', 'logo', 'taylor', 'swift', 'taylor', 'taylor', 'swift', 'performing', 'taylor', 'swift', 'tiktok', 'taylorswiftandtraviskelce', 'candace', 'owens', 'taylor', 'swift', 'margaret', 'qualley', 'taylor', 'swift', 'taylorswift']

References

[1] Hamid Abdollahpouri, Masoud Mansoury, Robin Burke, Bamshad Mobasher, and Edward Malthouse. User-centered evaluation of popularity bias in recommender systems. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, Utrecht, Netherlands, 2021.

[2] Anna Bergström and Maria J Belfrage. Incidental consumption and the role of opinion leaders. *News in Social Media*, 6(5):583–598, 2018.

[3] Maximilian Boeker and Alon Urman. An empirical investigation of personalization factors on tiktok. In *Proceedings of the ACM Web Conference 2022*, pages 2298–2309. Association for Computing Machinery, 2022.

[4] Luca Cervi and Clara Marín Lladó. Freepalestine on tiktok: From performative activism to (meaningful) playful activism. *Journal of International and Intercultural Communication*, 15(4):414–434, 2022.

[5] Albert Damstra, Rens Vliegenthart, Hajo Boomgaarden, Katharina Glüer, Erik Lindgren, Jesper Strömbäck, and Yariv Tsfati. Knowledge and the news: An investigation of the relation between news use, news avoidance, and the presence of (mis)beliefs. *The International Journal of Press/Politics*, 28(1):29–48, 2023.

[6] dfreelon. pyktok. `https://github.com/dfreelon/pyktok`, 2024.

[7] Noura Hagar and Nicholas Diakopoulos. Algorithmic indifference: The dearth of news recommendations on tiktok. *New Media & Society*, pages 1–21, 2023.

[8] Daniel Klug, Yiqing Qin, Michael Evans, and Geoffrey Kaufman. Trick and please: A mixed-method study on user assumptions about the tiktok algorithm. In *Proceedings of the 13th ACM Web Science Conference*, pages 84–92. Association for Computing Machinery, 2021.

[9] K. E. Matsa. More americans are getting news on tiktok, bucking the trend seen on most other social media sites, November 15 2023.

[10] OpenAI. Whisper: Towards Self-supervised Speech Recognition. `https://github.com/openai/whisper`, 2024.

[11] Markus V Reiss. Dissecting non-use of online news: Systematic evidence from combining tracking and automated text classification. *Digital Journalism*, 11(2):363–383, 2023.

[12] Kjerstin Thorson. Attracting the news: Algorithms, platforms, and reframing incidental exposure. *Sage Journals*, 21(8):1067–1082, 2020.

[13] Moran Yarchi and Liron Boxman-Shabtai. The image war moves to tiktok: Evidence from the may 2021 round of the israeli-palestinian conflict. *Digital Journalism*, 2023.