Guiding Sentiment Analysis with Hierarchical Text Clustering: Analyzing the German X/Twitter Discourse on Face Masks in the 2020 COVID-19 Pandemic

Silvan Wehrli[1] Chisom Ezekannagha[1] Georges Hattab[1,2]
T. Sonia Boender[3,4,5] Bert Arnrich[6] Christopher Irrgang[1]
[1]Centre for Artificial Intelligence in Public Health Research (ZKI-PH),
Robert Koch Institute, Berlin, Germany
{WehrliS,EzekannaghaC,HattabG,IrrgangC}@rki.de
[2]Department of Mathematics and Computer Science, Freie Universität, Berlin, Germany
[3]Department of Infectious Diseases, Public Health Service Amsterdam,
Amsterdam, Netherlands
sboender@ggd.amsterdam.nl
[4]Department of Health Sciences, Faculty of Science, Amsterdam Public Health Research
Institute Amsterdam and Amsterdam Institute for Immunology and
Infectious Diseases, Vrije Universiteit, Amsterdam, Netherlands
[5]Risk Communication Unit, Robert Koch Institute, Berlin, Germany
[6]Digital Health - Connected Healthcare, Hasso Plattner Institute,
University of Potsdam, Germany
Bert.Arnrich@hpi.de

Abstract

Social media are a critical component of the information ecosystem during public health crises. Understanding the public discourse is essential for effective communication and misinformation mitigation. Computational methods can aid these efforts through online social listening. We combined hierarchical text clustering and sentiment analysis to examine the face mask-wearing discourse in Germany during the COVID-19 pandemic using a dataset of 353,420 German X (formerly Twitter) posts from 2020. For sentiment analysis, we annotated a subsample of the data to train a neural network for classifying the sentiments of posts (neutral, negative, or positive). In combination with clustering, this approach uncovered sentiment patterns of different topics and their subtopics, reflecting the online public response to mask mandates in Germany. We show that our approach can be used to examine long-term narratives and sentiment dynamics and to identify specific topics that explain peaks of interest in the social media discourse.

1 Introduction

Social media platforms play an essential role in the information ecosystem during public health emergencies such as disease outbreaks, as they are widely used (We Are Social et al., 2024a) and catalyze the dissemination of information (Vraga et al., 2023). The public turns to these platforms to look for information, share and access news, express opinions, and exchange personal experiences (We Are Social et al., 2024b). When there is an overabundance of information available during health emergencies, this is called an *infodemic* (Briand et al., 2023). Infodemics may include any information, accurate or false, i.e., misinformation, regardless of the intention (Lewandowsky et al., 2020). Understanding the information ecosystem of infodemics is crucial for developing effective data-driven and human-centered public health communication that addresses concerns and mitigates harmful effects from misinformation (Borges do Nascimento et al., 2022; Briand et al., 2023) and for infodemic preparedness (Wilhelm et al., 2023).

In the context of social media, natural language processing can help to monitor the public discourse (Baclic et al., 2020). This monitoring is commonly referred to as *social listening* (Stewart and Arnold, 2018), a key research field in infodemic management (Calleja et al., 2021). While it is often used in digital marketing, social listening is relatively new to the public health domain (Boender et al., 2023). In social listening, the classification of social media data into topics is used to identify different aspects of online conversations (*topic analysis*) and to measure temporal relevance over time (Purnat et al., 2021). To this end, the World Health Organization's Early Artificial Intelligence–Supported Response With Social Listening Platform (EARS, White et al. (2023)) used semi-supervised machine learning for classifying social media content into

Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 153–167
August 15, 2024 ©2024 Association for Computational Linguistics

topics, which offered real-time analytics to public health researchers during the COVID-19 pandemic. Other works have used unsupervised methods, in particular, *topic modeling* (Blei, 2012), which represents topics as word distributions through generative probabilistic modeling (e.g., Rowe et al. (2021)), and *text clustering* (Willett, 1988), which represents topics as groups of semantically similar texts (e.g., Santoro et al. (2023)). In addition, *sentiment analysis* (Liu, 2012) can improve the understanding of the public perception of health-related topics by classifying sentiments expressed in texts (Boender et al., 2023; Briand et al., 2023).

While some studies have combined these techniques (e.g., Rowe et al. (2021)), they have typically used a *flat* representation of the data, i.e., a fixed number of topics in one level. In contrast, text data can be represented *hierarchically* on multiple levels, i.e., subgroups within one topic, with varying cluster sizes and granularities (Aggarwal and Zhai, 2012). The representation as a hierarchy allows the structured exploration of large document collections (Cutting et al., 1992) and helps to identify online narratives on social media in the context of public health (White et al., 2023).

In this work, we combine sentiment analysis with hierarchical text clustering to analyze a German X (formerly Twitter) dataset on wearing face masks during the COVID-19 pandemic in 2020. In Germany, the mask requirement was introduced at the end of April 2020 for public transport and stores (Die Bundesregierung, 2020b). The introduction of the obligation was preceded by a lockdown from mid-March with contact restrictions and the closure of numerous facilities in public spaces, e.g., schools as a consequence of an increase in COVID-19 cases (Die Bundesregierung, 2020c). The first easing of restrictions was implemented in mid-April (MDR, 2020). In order to extract sentiments from this much debated time period, we annotated a subsample of the selected dataset for sentiment analysis and trained a neural network for sentiment classification. We analyze the combined results in the context of the COVID-19 pandemic in Germany. Based on the overview of high-level coarse clusters and corresponding sentiments, we identify topics of interest for an in-depth analysis. We demonstrate the ability of our approach to systematically analyze highly debated public health measures such as face masks (Deutschlandfunk, 2020; MDR, 2020), which significantly impacted daily life in Germany.

2 Related Work

In the following, we discuss related work, focusing on machine learning techniques and applications relevant to X data and the German language.

2.1 Sentiment Analysis

Regarding German language sentiment analysis, machine learning methods typically outperform lexicon-based methods, and neural network models typically outperform traditional machine learning (Borst et al., 2023; Schmidt et al., 2022; Struß et al., 2019; Zielinski et al., 2023).

Guhr et al. (2020) fine-tuned a neural network for classification using a broad range of German sentiment datasets (GBERT$_{\text{broad}}$), including two datasets with X posts. GBERT$_{\text{broad}}$ builds on GBERT (Chan et al., 2020), a BERT transformer-based encoder (Devlin et al., 2019) pretrained exclusively on German text. GBERT is also used successfully for fine-tuning sentiment classifiers on other task-annotated data (e.g., Schmidt et al. (2022); Zielinski et al. (2023)).

XLM-T (Barbieri et al., 2022) is a multilingual sentiment classifier for X posts trained on eight languages, including German. It is based on the multilingual XLM-RoBERTa (Conneau et al., 2020), which also uses the BERT architecture. Notably, it benefits from additional pretraining on posts prior to supervised fine-tuning, which may improve the performance on supervised classification tasks (Gururangan et al., 2020).

2.2 Text Clustering

Xu et al. (2015) suggest that *embeddings*, i.e., high-dimensional vector representations derived through language modeling (Vinokourov et al., 2002), yield better results as inputs for text clustering than the traditionally used *bag of words*, i.e., numeric representations based on word occurrences (Aggarwal and Zhai, 2012).

Embedding-based text clustering is proposed as an alternative to topic modeling for identifying topics in text data (e.g., Angelov (2020); Sia et al. (2020)). Unlike topic modeling, text clustering does not assign descriptive keywords to topics. These need to be extracted separately using techniques like *term frequency–inverse document frequency* (TF-IDF). This statistical measure calculates the relevance of words in a text collection (Ramos, 2003). We use embedding-based text clustering since it can be advantageous for social media data, as it may work better with short texts (Egger and Yu, 2022).

Creating hierarchies for text collections, as opposed to flat clustering, can help to explore and understand the contextual relationships (Cutting et al., 1992). Hierarchical clustering algorithms are often computationally expensive (Aggarwal and Zhai, 2012), limiting their use on large datasets. In this work, we use the Sub-Cluster Component Algorithm by Monath et al. (2021), who address this issue through various conceptional improvements compared to the traditional hierarchical agglomerative clustering without sacrificing clustering quality. Through the use of this algorithm, the clustering is based entirely on text embeddings. This is in contrast to the text clustering framework BERTopic (Grootendorst, 2022), which enables hierarchical text clustering, but combines the clustering of text embeddings and bag of words.

2.3 COVID-19-specific X Analysis

Various studies used sentiment analysis to analyze the online debate on X around COVID-19 during the pandemic in Germany. Reiter-Haas et al. (2023) analyzed the debate on contact tracing, vaccination, and face masks and contrasted the results with survey results. Schmidt et al. (2022) focused on the 2021 federal election in Germany. They analyzed the change in sentiments of the political parties' posts in the election. Rowe et al. (2021) used topic modeling and sentiment analysis to analyze X data from Germany, other European countries, and the United States to understand the sentiment towards immigration during the early stage of the COVID-19 pandemic in 2020. None of these studies used text clustering.

Santoro et al. (2023) used flat text clustering to analyze the different aspects of the online debate about vaccination in different countries over time. However, they did not consider sentiments.

In non-German analyses, Sanders et al. (2021) combined text clustering and sentiment analysis to study English face mask-related posts. They represented topics in a two-level hierarchy. Purnat et al. (2021) developed a more fine-grained hierarchy of five levels for classifying COVID-19 online conversations in English and French. However, posts were classified into topics with manually defined keywords. This taxonomy then served as the basis for the semi-supervised topic classification in EARS (White et al., 2023).

This work combines and expands on the ideas of Sanders et al. (2021) and Purnat et al. (2021) and presents a social listening approach for public health that unifies topic and sentiment analysis. Our approach allows a flexible representation of the hierarchy with an adjustable number of levels. Additionally, our work contributes to the analysis of the social media discourse during the COVID-19 pandemic in Germany, surpassing the time period of data considered in previous work (Reiter-Haas et al., 2023).

3 Data

In this section, we describe the collection of X data and the dataset construction for sentiment analysis.

3.1 X (formerly Twitter) Data

Collecting German posts Between November 2022 and April 2023, we collected 50% of all original posts (i.e., excluding replies, comments, or quotes) in the German language for the year 2020 using the Academic Research API (X, 2023a). We used the post counts API (X, 2024) to estimate the number of original posts in 2020 per minute. For each minute (the smallest possible sampling time period for API queries), the API returned 50% of posts starting from the end of each minute. As true random sampling is not possible with the research API, we used the smallest possible time period, aiming to retrieve a representative sample of the entire stream that reflects its temporal characteristics, e.g., day/night shifts, and discussed topics for posts in the German language. We used the language tag provided by the API and refined the data through language identification with FastText (Joulin et al., 2017). We limited the posts to 2020 because X terminated our API access in April 2023. The final dataset contained 38 million posts.

Face mask dataset We filtered posts by words used to describe face masks: 'Maske' ('mask'), 'Mundschutz' and 'Mund-Nasen-Schutz' ('face mask' or 'surgical mask'), and 'FFP2', resulting in a *face mask dataset* with 353,420 posts.[1] Using a sample of 1,000 included and excluded posts, we calculated a precision of 97.4% and a recall of 100% (cf. Limitations). We note that this dataset is not limited to posts originating from Germany but includes any posts in German. In line with

[1] For 'Mund-Nasen-Schutz', we included 'Mundnasenschutz' and 'Mund-Nasenschutz' as variations. We excluded posts containing '#maskedsinger', '#themaskedsinger', 'maskedsinger', or 'masked singer' (relating to a German TV show). We used lowercasing for filtering.

our Ethics Statement, we exclude any location data from our analysis.

For context, we included the *7-day incidence rate*, i.e., the sum of COVID-19 cases in Germany with a reporting date within the last seven days, based on 100,000 inhabitants (Robert Koch-Institut, 2024).

3.2 Sentiment Analysis Data

Data selection We sampled 2,200 posts from the face mask dataset, using weights based on hourly relative frequency to maintain a similar temporal distribution. We adjusted the weights using the square root of relative frequencies to avoid oversampling periods with high post volume. This ensured that events related to the face mask requirement in Germany, which may have led to an increased post volume in the face mask dataset, were proportionally represented in the sample.

Annotation We asked seven annotators to label the general sentiments expressed in posts. Every post was labeled by three different annotators (at least 500 posts per annotator). Annotators were instructed to classify posts into four distinct categories: *neutral*, *negative*, *positive*, and *mixed* (containing *negative* and *positive* sentiments). We provided annotators with instructions and examples (not included in the final dataset). We used the majority rule to decide on the final label of posts (Table 1) and excluded samples without a majority from the final dataset. To measure the agreement between annotators, we report a Fleiss' κ (Fleiss and Cohen, 1973) of 0.60, calculated using statsmodels (Seabold and Perktold, 2010). This score indicates moderate agreement (Landis and Koch, 1977) and is comparable to Schmidt et al. (2022), who annotated X posts by German politicians in a similar setting.

Sentiment	Count	Percentage
neutral	876	40.26%
negative	858	39.45%
positive	239	10.99%
mixed	130	5.98%
no majority	72	3.31%

Table 1: Results of the data annotation for sentiment analysis based on samples from the face mask dataset.

Data splitting Finally, we split the annotated dataset into training, validation, and test sets using a 7:1:2 stratified random split, i.e., maintaining the original class distribution (Table 1) across splits. We only used posts with neutral, negative, or positive labels. We removed mixed posts to establish a stronger baseline for distinguishing between the primary sentiment classes.

4 Methodology

In the follwoing, we outline the training of the sentiment classifier, the application of hierarchical clustering, and the analytical approach.

4.1 Sentiment Analysis

For training the sentiment classifier, the base version of GBERT (Chan et al., 2020) serves as a starting point. GBERT has shown competitive results in German sentiment analysis (Schmidt et al., 2022; Zielinski et al., 2023). We consider two scenarios: First, using GBERT out-of-the-box for initializing a classifier. Second, we continue pretraining on the face mask dataset using whole word 'masking' similar to GBERT. This excludes the sentiment analysis data. For continued pretraining, we use the hyper-parameter setup for task-adaptive pretraining (TAPT) as suggested by Gururangan et al. (2020). For the supervised fine-tuning, we use the hyper-parameter setup suggested by Devlin et al. (2019). We base model selection on validation set performance, with an evaluation carried out every 10 steps. The training is performed on a single NVIDIA A100 GPU with the PyTorch (Ansel et al., 2024) and Transformers (Wolf et al., 2020) frameworks.

4.2 Hierarchical Clustering

The Sub-Cluster Component Algorithm (SCC, Monath et al. (2021)) is used for hierarchical text clustering thanks to its competitive performance on large datasets. SCC operates on nearest-neighbor similarity, repeatedly merging clusters to build a tree with multiple partition levels. The user controls these levels and the minimum similarity threshold for cluster merging. In this work, we represent posts as embeddings, using cosine similarity to define similarity. Cosine similarity is a standard metric to measure semantic similarity for vector-based text representations (Chandrasekaran and Mago, 2021) and is used by Monath et al. (2021).[2]

In the first step, posts are embedded using *German BERT large paraphrase cosine* (May et al.,

[2] We provide code on GitHub: https://github.com/ClimSocAna/sentiments-with-hierarchical-clustering.

2023), a GBERT-large model fine-tuned for representing text similarity using cosine similarity. GBERT models show competitive results for German language text clustering (Wehrli et al., 2023), making *German BERT large paraphrase cosine* well-suited for the use with the SCC algorithm.

In the second step, the dataset is transformed into a nearest neighbors graph using Faiss (Douze et al., 2024). This graph is then used as an input to the SCC algorithm, with the state-of-the-art parameter setup (Monath et al., 2021), which includes 200 rounds of geometrically increasing thresholds and average linkage clustering. To improve computational efficiency, Monath et al. (2021) use a highly sparsified nearest neighbors graph (considering the 25 closest neighbors) to approximate cluster similarity. However, in this work, the number of neighbors considered is increased as much as possible (10,000 neighbors) to obtain the most accurate clustering possible. Given our computational resources, this results in RAM use of roughly 300 GB and a runtime of under 10 hours in a multi-CPU setup.

4.3 Analytical Approach

Topic size To measure a topic's importance, we use the number of posts it contains to represent the range of discussed content. Depending on the research question, however, this metric could be adapted, for example, by including the popularity of posts to give greater weight to social resonance.

Sentiment score We use a *sentiment score* to analyze the sentiment of topics. This score is defined as

$$score_{t,p} = \frac{|posts_{t,p,pos}| - |posts_{t,p,neg}|}{|posts_{t,p,all}|}, \quad (1)$$

where $posts$ denotes the set of posts for a topic t, a period of time p, and for specific sentiments (neg for negative, pos for positive, all for all categories). The sentiment score calculates the average sentiment for a set of posts, considering positive and negative posts as polar values (+1, respectively, -1). We use this metric to highlight differences in topics based on their sentiment composition.

Cluster selection The SCC's multi-level output allows us to analyze topics in varying detail. To demonstrate the flexibility of the hierarchical clustering, we select three levels of increasing topic granularity with 20, 104, and 1,051 clusters, respectively.

Cluster labeling and validation We first extract descriptive keywords for each cluster using class-based TF-IDF (Grootendorst, 2022). This is a variation of the traditional TF-IDF, which emphasizes the distinctiveness of keywords between clusters. We limit the set of candidate words to lemmatized content words to increase the information value of keywords, using tokenization (Proisl and Uhrig, 2016), part-of-speech tagging (Proisl, 2018) and lemmatization (Schmid, 1999). We select components optimized for German social media text based on Ortmann et al. (2019).[3] We validate clustering quality through keyword analysis and random sampling of 200 posts per cluster to ensure the analyzed clusters represent distinct topics. The selection process could be supplemented by measures that quantify the quality of individual levels through intra- and inter-cluster (dis)similarity. We manually extract labels for each analyzed cluster based on the extracted keywords and the sampled posts.

Cluster visualization Hierarchical clustering organizes data into a tree-like structure called a dendrogram, where each node represents a cluster, and the branches show the relationships between them. The dendrogram's structure is often used to visualize the hierarchical relationships between the clusters. It is important to note that the clusters are, fundamentally, sets of data points, not tree structures. However, visualizations based on dendrograms are limited by the quantification of the underlying clustering metrics and by their lack of flexibility and customizability.

Given these limitations, we use a treemap idiom to provide a compact and intuitive way (Hattab et al., 2020) to navigate and explore the resulting hierarchies. The treemaps visualize the hierarchical relationships between clusters by making the size of each node proportional to the relative importance or size of the cluster. The hierarchical structure is conveyed by the nested layout of the treemap, where child nodes are contained within their parent nodes. This allows the visualization of complex topic hierarchies and identifies dominant topics and their relationships.

Two use cases are considered to illustrate the treemap idiom. They address the relationships of subclusters to their parent cluster and the temporal

[3] We provide a spaCy-based (Honnibal et al., 2020) implementation on GitHub: `https://github.com/slvnwhrl/GerSoMeTokenExtractor`.

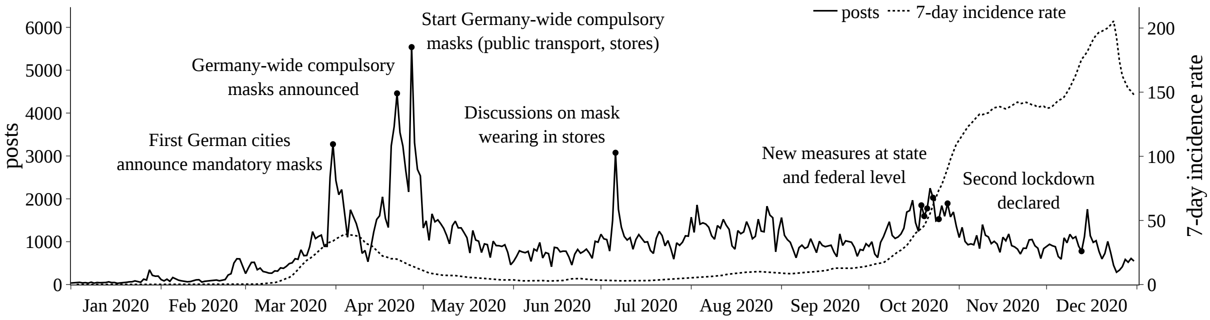

Figure 1: The number of German X posts discussing *face masks* relative to the 7-day incidence rate of COVID-19 cases (Robert Koch-Institut, 2024) and selected events (DW, 2020; MDR, 2020; Tagesschau, 2020a,b) in Germany.

changes of the relative importance of one subcluster. This relies on the overview detail and the small multiple idioms (Shneiderman, 2003) and corresponds to Figures 2 and 3, respectively.

5 Results & Analysis

In this section, we evaluate the results of the sentiment classification training, followed by the analysis of the face mask dataset.

5.1 Evaluation of the Sentiment Classification

Results Table 2 reports the overall results for the classification of neutral, negative, and positive sentiments of posts on the test set for the fine-tuned GBERT models (as outlined in Subsection 4.1). Additionally, we evaluated GBERT$_{\text{broad}}$ and XLM-T as baselines. GBERT$_{\text{SFT}}$, fine-tuned on the annotated face mask posts, achieved an average weighted F1-score of 77.90%. GBERT$_{\text{broad}}$ fell significantly behind with a more than 20 percentage points lower F1$_{\text{weighted}}$. The performance gap of XLM-T is much smaller at less than four percentage points. GBERT$_{\text{TAPT+SFT}}$ delivered the best results, achieving an 3.06 percentage points higher average weighted F1-score than GBERT$_{\text{SFT}}$ through additional pretraining. Based on these results, we used the best-performing GBERT$_{\text{TAPT+SFT}}$ to analyze the face mask dataset.[4]

Error analysis The most common errors of GBERT$_{\text{TAPT+SFT}}$ were neutral posts misclassified as negative, vice versa, and positive posts misclassified as neutral (cf. Table 3, Appendix A). Predictions for the neutral and negative classes showed relatively balanced precision and recall compared to the positive class (cf. Table 4, Appendix A). The lower overall F1-score (68.97%) and comparatively lower recall (62.50%) of the positive class is likely, to some degree, a result of the class imbalance (Table 1, Johnson and Khoshgoftaar (2019)).

The inspection of misclassified samples showed that the model sometimes struggled with implicit sentiment and sarcasm, likely contributing to the lower performance of the positive class (Riloff et al., 2013).

Model	Accuracy	F1$_{\text{macro}}$	F1$_{\text{weighted}}$
GBERT$_{\text{SFT}}$	79.24%	75.77%	79.45%
	78.08%	*73.49%*	*77.90%*
GBERT$_{\text{TAPT+SFT}}$	82.53%	79.13%	82.36%
	81.06%	*77.60%*	*80.96%*
GBERT$_{\text{broad}}$	56.20%	47.57%	54.05%
XLM-T	74.18%	71.24%	74.20%

Table 2: Test set results for sentiment classification of face mask-related X posts. For GBERT$_{\text{SFT}}$ and GBERT$_{\text{TAPT+SFT}}$, we report single-best model results and the average of five models with different seeds (in *italic*). GBERT$_{\text{broad}}$ (Guhr et al., 2020) and XLM-T (Barbieri et al., 2022) are models from the literature, serving as baselines.

5.2 Results and Discussion of the Face Mask Dataset

The face mask dataset represents the dynamic social media discourse about face masks during the COVID-19 pandemic (Figure 1). The rise in notified COVID-19 cases, i.e., the 7-day incidence rate, and introductions of public health interventions were often associated with increased online conversation on this topic. Some events stand out, especially the initial introduction of the mask requirement in April 2020 (Figure 1). The falling number of cases in the summer of 2020 fueled the debate about the necessity of wearing face masks, for example, at the beginning of July. With the ris-

[4] We released the best-performing model on Hugging Face: `https://huggingface.co/slvnwhrl/gbert-face-mask-sentiment`.

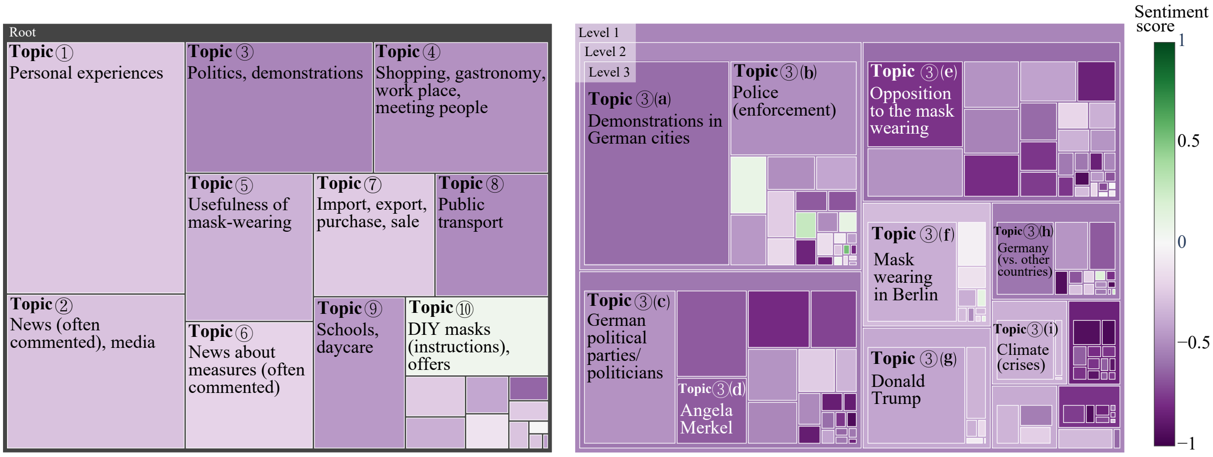

Figure 2: *Left*: Treemap depicting topics of German language face mask-related discourse on X in 2020. Posts are hierarchically clustered into 20 topics. Cluster sizes reflect the number of posts in each topic; colors encode the value of the sentiment score (Equation 1), which represents the mean sentiment of posts for each topic (1 = completely positive, -1 = completely negative). GBERT$_{\text{TAPT+SFT}}$ was used to assign sentiments to posts. Topic labels were manually extracted from 200 randomly sampled posts per cluster and shown for the 10 largest topics. *Right*: Treemap depicting the subtopics of topic ③ (level 1) on the two lower levels of the clustering hierarchy (level 2 and 3). The largest subtopics on level 3 were manually labeled.

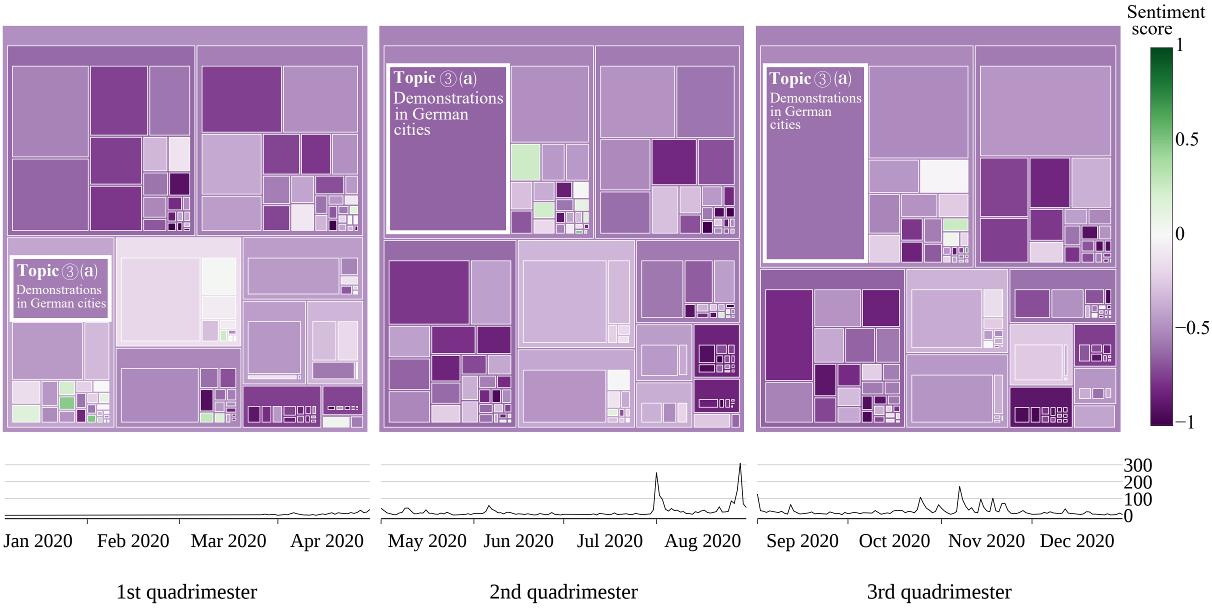

Figure 3: *Top*: Small multiples of the treemap of topic ③ (Figure 2, *right*) for the three quadrimesters of 2020. Subtopic ③(a) ('Demonstrations in German cities') is highlighted through a thicker border. The treemaps show the change in relative size and sentiment score of the subtopics over time. Representation of cluster sizes and colors follow the same approach as in Figure 2. *Bottom*: Small multiples of frequency charts showing the number of posts in subtopic ③(a), allowing to identify times with high post volume.

ing case numbers in autumn, new public health and social measures were implemented, which included the restriction of social contacts and the temporary closure of restaurants or cultural institutions, but less restrictive than during the first lockdown in March 2020 (Die Bundesregierung, 2020a,c). Finally, the federal government imposed a second lockdown before Christmas (Die Bundesregierung, 2020d). These dates showed a lower number of posts than during the first introduction of mandatory face mask wearing.

Topics overview The dynamic nature of the debate can also be seen in the overarching topics of the discourse on X, resulting from the hierarchical clustering (Figure 2). We provide an interactive visualization on GitHub.[5] The most prominently discussed topics are areas of life that were often the target of public health measures, namely stores (④), public transport (⑧), or schools (⑨). When reviewing samples, we observed that many posts contain descriptions of everyday life, such as other people's behavior when shopping, e.g., when neglecting the face mask mandate. The data show that a large part of posts are about sharing experiences (①) and include a large variety of topics.

Sharing news also plays an important role in the COVID-19 discourse on face masks; on the one hand, news specific to the introduction of the mask requirement (②), but also more general news on the subject of masks, such as reports on the number of COVID-19 cases (⑥). It is noticeable that posts with shared news (e.g., with a link) often also contain the user's opinion about the content. We identified similar tendencies for topics that mainly revolve around the benefits of masks (⑤) and the political and social debate surrounding the obligation to wear masks (③).

Finally, some topics deal primarily with the production of do-it-yourself (DIY) masks (⑩) and the procurement of masks (⑦).

Overall, posts expressing negative sentiment dominated the discourse (Figure 2, *left*). In fact, more negative than positive posts were published at all times during the investigated year (cf. Figure 4, Appendix A). Some topics are particularly negative (③, ④, ⑧, ⑨), with the topic on DIY masks (⑩) standing out as the only topic with a slightly positive sentiment. During the analysis, we often noticed posts in which users shared their DIY masks. In Figure 2, we show the subtopics, i.e., deeper level, of the topic of politics and demonstrations (③), to present a more differentiated view of this particularly negative topic. Within this topic, users discuss German or international politicians (e.g., Angela Merkel (③(d)) or Donald Trump (③(g))), Germany in international comparison (③(h)), or topics related explicitly to Berlin (③(f)) or the climate (③(i)). This view reveals differences in the relative importance and sentiment of subtopics that make up the overall negative sentiment of the higher-level topic and shows that mask wearing is discussed in very different political contexts.

Demonstrations The topic of demonstrations in German cities is particularly striking due to its size and negativity (Figure 2, *right*). Thus, we chose this topic for a more detailed analysis, exemplifying the capabilities of hierarchical text clustering to guide sentiment analysis. In Germany, the first demonstrations against public measures took place at the beginning of May, with larger demonstrations occurring repeatedly throughout the year, for example, in Berlin and Leipzig (MDR, 2020). Figure 3 shows the size and sentiment of the subtopic cluster in the parent topic of politics and demonstrations (*top*) and the number of posts on demonstrations in German cities (*bottom*) in the three quadrimesters of 2020. We found that demonstrations are more frequently discussed over the course of 2020 and contribute more to the parent topic without significant changes in negativity (Figure 3, *top*). Pointwise increases in posts occurred at the time of specific demonstrations (Figure 3, *bottom*), e.g., two large demonstrations with tens of thousands of participants in Berlin at the beginning and end of August 2020 (MDR, 2020).

Comparing results Reiter-Haas et al. (2023) also investigated the sentiments of face mask-related German posts from X. They did not find a clear tendency towards positive or negative sentiments. Compared to our study, they used a smaller X sample for sentiment analysis (15,425 versus 353,420 posts), only considered data from January to August 2020 (as opposed to data from the whole year), and used a different method for sentiment analysis (based on a sentiment lexicon). These factors may explain the different outcome of our study, as we find that the majority of posts are negative.

Sanders et al. (2021) used two-level hierarchical text clustering to analyze English X posts on face masks from March to July 2020. They identified topic clusters of varying sentiments, similar to our findings. However, posts with negative sentiments did not dominate the overall discourse. Additionally, topical parallels can be drawn between the discourse on political actors (such as Donald Trump) or on public measures such as the requirement to wear face masks in stores, and the sharing of personal experiences from everyday life. Similar to Reiter-Haas et al. (2023), they used lexicon-based sentiment analysis, albeit on a larger sample (1,013,039 posts).

6 Conclusion and Outlook

This study employed hierarchical text clustering and sentiment analysis to examine the public dis-

[5] `https://github.com/ClimSocAna/sentiments-with-hierarchical-clustering`.

course surrounding face masks in Germany during the COVID-19 pandemic 2020. Our analysis of 353,420 face mask-related posts reveals the dynamics of the German public's online response on X (formerly Twitter) to mask mandates. Our findings indicate an overall negative sentiment dominating face mask-related posts. Analyzing specific topics revealed nuanced sentiment patterns. For instance, the topic of DIY masks was slightly positive, while topics linked to COVID-19 demonstrations and general political discourse on face mask policies showed stronger negativity. We show that the combination of clustering, sentiment classification, and suitable visualization helps analyze complex social media discourse in a structured manner. This study thereby advances the methods of social listening in public health. Furthermore, our analysis contributes to the understanding of the online information ecosystem during the COVID-19 pandemic in Germany. This is a prerequisite to better understand the (harmful) impact of the infodemic during the COVID-19 pandemic on the public. Ultimately, this enables knowledge-based preparation for a future pandemic that will likely be accompanied by an infodemic again (Briand et al., 2023; Wilhelm et al., 2023). In this context, the proposed approach enables structured analyses of social media data for topics that are relevant for the review of the COVID-19 pandemic, respectively, infodemic (such as mental health).

Our results offer a starting point for further research. The presented approach could be adapted for real-time infodemic surveillance ('infoveillance') to track sentiment dynamics and emerging topics during future health-related social media discourses, e.g., based on the online version of the SCC algorithm (Monath et al., 2023). To operationalize this approach during the next pandemic or public health crisis, developing a framework for interactive data exploration is required to generate insights that can inform public health action. Finally, specific investigations into the role of misinformation within negative clusters are needed to illuminate public health communication challenges.

Limitations

Keyword filtering We noticed that a few posts were filtered incorrectly as relevant because 'Maske' ('mask') was used as a homonym for a beauty product or as part of a costume. Furthermore, other terms may be relevant to the online discussion about face masks, even if our analysis showed a high recall. For example, there are colloquial or dialect words to consider, such as 'Schnutenpulli' (NORD24, 2020), originally from Low German. Finally, we did not consider misspellings.

Clustering While the chosen clustering algorithm and language model have proven effective, a key limitation is their singularity. Exploring multiple clustering algorithms and language models could reveal different data structures and provide more nuanced insights. Comparing the results from diverse methods would help assess the robustness of the observed clusters and sentiment patterns.

Social media data Finally, we note that accessing data from social media platforms remains challenging, which currently limits the application of our approach to X data. The European Union's *Digital Services Act* is likely to improve this situation for infodemic research and practice in the future, as it will legally enable researchers to access data on large platforms (Wehrli et al., 2024).

Ethics Statement

The X (formerly Twitter) data collected as part of this work is subject to X's Developer Agreement and Policy (X, 2023b) and the European Union's General Data Protection Regulation (GDPR, European Commission), which we comply with. We only process post texts and timestamps, remove user mentions and URLs from the post texts, and do not use any post metadata that allows the identification of individuals (such as user names or location data). In addition, we only present results at an aggregated level, i.e., for groups of posts. We do not publish or share any of the data or results in a way that does not align with X's Developer Agreement and Policy and the GDPR.

Acknowledgments

We would like to thank the annotators of the sentiment analysis dataset for their help in realizing this work. We would also like to thank the reviewers for their time and valuable input.

References

Charu C. Aggarwal and ChengXiang Zhai. 2012. *A Survey of Text Clustering Algorithms*, pages 77–128. Springer US, Boston, MA.

Dimo Angelov. 2020. Top2Vec: Distributed representations of topics. *Preprint*, arXiv:2008.09470.

Jason Ansel, Edward Yang, Horace He, Natalia Gimelshein, Anjali Jain, Michael Voznesensky, Bin Bao, Peter Bell, David Berard, Evgeni Burovski, Geeta Chauhan, Anjali Chourdia, Will Constable, Alban Desmaison, Zachary DeVito, Elias Ellison, Will Feng, Jiong Gong, Michael Gschwind, Brian Hirsh, Sherlock Huang, Kshiteej Kalambarkar, Laurent Kirsch, Michael Lazos, Mario Lezcano, Yanbo Liang, Jason Liang, Yinghai Lu, C. K. Luk, Bert Maher, Yunjie Pan, Christian Puhrsch, Matthias Reso, Mark Saroufim, Marcos Yukio Siraichi, Helen Suk, Shunting Zhang, Michael Suo, Phil Tillet, Xu Zhao, Eikan Wang, Keren Zhou, Richard Zou, Xiaodong Wang, Ajit Mathews, William Wen, Gregory Chanan, Peng Wu, and Soumith Chintala. 2024. PyTorch 2: Faster machine learning through dynamic python bytecode transformation and graph compilation. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, ASPLOS '24, page 929–947, New York, NY, USA. Association for Computing Machinery.

Oliver Baclic, Matthew Tunis, Kelsey Young, Coraline Doan, Howard Swerdfeger, and Justin Schonfeld. 2020. Challenges and opportunities for public health made possible by advances in natural language processing. *Can Commun Dis Rep*, 46(6):161–168.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266, Marseille, France. European Language Resources Association.

David M. Blei. 2012. Probabilistic topic models. *Commun. ACM*, 55(4):77–84.

T Sonia Boender, Paula Helene Schneider, Claudia Houareau, Silvan Wehrli, Tina D Purnat, Atsuyoshi Ishizumi, Elisabeth Wilhelm, Christopher Voegeli, Lothar H Wieler, and Christina Leuker. 2023. Establishing infodemic management in Germany: A framework for social listening and integrated analysis to report infodemic insights at the national public health institute. *JMIR Infodemiology*, 3:e43646.

Israel Júnior Borges do Nascimento, Ana Beatriz Pizarro, Jussara Almeida, Natasha Azzopardi-Muscat, Marcos André Gonçalves, Maria Björklund, and David Novillo-Ortiz. 2022. Infodemics and health misinformation: a systematic review of reviews. *Bull. World Health Organ.*, 100(9):544–561.

Janos Borst, Jannis Klaehn, and Manuel Burghardt. 2023. Death of the dictionary? - the rise of zero-shot sentiment classification. In *Proceedings of the Computational Humanities Research Conference 2023, Paris, France, December 6-8, 2023*, volume 3558 of *CEUR Workshop Proceedings*, pages 303–319. CEUR-WS.org.

Sylvie Briand, Sarah Hess, Tim Nguyen, and Tina D. Purnat. 2023. *Infodemic Management in the Twenty-First Century*, pages 1–16. Springer International Publishing, Cham.

Neville Calleja, AbdelHalim AbdAllah, Neetu Abad, Naglaa Ahmed, Dolores Albarracin, Elena Altieri, Julienne N Anoko, Ruben Arcos, Arina Anis Azlan, Judit Bayer, Anja Bechmann, Supriya Bezbaruah, Sylvie C Briand, Ian Brooks, Lucie M Bucci, Stefano Burzo, Christine Czerniak, Manlio De Domenico, Adam G Dunn, Ullrich K H Ecker, Laura Espinosa, Camille Francois, Kacper Gradon, Anatoliy Gruzd, Beste Sultan Gülgün, Rustam Haydarov, Cherstyn Hurley, Santi Indra Astuti, Atsuyoshi Ishizumi, Neil Johnson, Dylan Johnson Restrepo, Masato Kajimoto, Aybüke Koyuncu, Shibani Kulkarni, Jaya Lamichhane, Rosamund Lewis, Avichal Mahajan, Ahmed Mandil, Erin McAweeney, Melanie Messer, Wesley Moy, Patricia Ndumbi Ngamala, Tim Nguyen, Mark Nunn, Saad B Omer, Claudia Pagliari, Palak Patel, Lynette Phuong, Dimitri Prybylski, Arash Rashidian, Emily Rempel, Sara Rubinelli, PierLuigi Sacco, Anton Schneider, Kai Shu, Melanie Smith, Harry Sufehmi, Viroj Tangcharoensathien, Robert Terry, Naveen Thacker, Tom Trewinnard, Shannon Turner, Heidi Tworek, Saad Uakkas, Emily Vraga, Claire Wardle, Herman Wasserman, Elisabeth Wilhelm, Andrea Würz, Brian Yau, Lei Zhou, and Tina D Purnat. 2021. A public health research agenda for managing infodemics: Methods and results of the first WHO infodemiology conference. *JMIR Infodemiology*, 1(1):e30979.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Dhivya Chandrasekaran and Vijay Mago. 2021. Evolution of semantic similarity—a survey. *ACM Comput. Surv.*, 54(2).

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Douglass R. Cutting, David R. Karger, Jan O. Pedersen, and John W. Tukey. 1992. Scatter/gather: a cluster-based approach to browsing large document collections. In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '92, page 318–329, New York, NY, USA. Association for Computing Machinery.

Deutschlandfunk. 2020. Chronologie eines Schuljahrs in der Coronakrise. `https:`

//www.deutschlandfunk.de/rueckblick-2020-chronologie-eines-schuljahrs-in-der-100.html. Accessed: 2024-04-12.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Die Bundesregierung. 2020a. Diese Regeln gelten jetzt. https://www.bundesregierung.de/breg-de/themen/coronavirus/regelungen-ab-2-november-1806818. Accessed: 2024-05-02.

Die Bundesregierung. 2020b. Maskenpflicht in ganz Deutschland. https://www.bundesregierung.de/breg-de/themen/coronavirus/maskenpflicht-in-deutschland-1747318. Accessed: 2024-04-12.

Die Bundesregierung. 2020c. Telefonschaltkonferenz der Bundeskanzlerin mit den Regierungschefinnen und Regierungschefs der Länder am 15. April 2020. https://www.bundesregierung.de/breg-de/aktuelles/bund-laender-beschluss-1744224. Accessed: 2024-05-02.

Die Bundesregierung. 2020d. "Wir sind zum Handeln gezwungen". https://www.bundesregierung.de/breg-de/themen/coronavirus/merkel-beschluss-weihnachten-1827396. Accessed: 2024-05-02.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The Faiss library. *Preprint*, arXiv:2401.08281.

DW. 2020. Masken gegen Corona: Wie lange noch? https://www.dw.com/de/masken-gegen-corona-wie-lange-noch/a-54060107. Accessed: 2024-04-30.

Roman Egger and Joanne Yu. 2022. A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify Twitter posts. *Frontiers in Sociology*, 7.

European Commission. Data protection in the EU. https://commission.europa.eu/law/law-topic/data-protection/data-protection-eu_en. Accessed: 2024-05-12.

Joseph L. Fleiss and Jacob Cohen. 1973. The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability. *Educational and Psychological Measurement*, 33(3):613–619.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *Preprint*, arXiv:2203.05794.

Oliver Guhr, Anne-Kathrin Schumann, Frank Bahrmann, and Hans Joachim Böhme. 2020. Training a broad-coverage German sentiment classification model for dialog systems. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1627–1632, Marseille, France. European Language Resources Association.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Georges Hattab, Theresa-Marie Rhyne, and Dominik Heider. 2020. Ten simple rules to colorize biological data visualization. *PLOS Computational Biology*, 16(10):e1008259.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Justin M. Johnson and Taghi M. Khoshgoftaar. 2019. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):27.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Stephan Lewandowsky, John Cook, and Doug Lombardi. 2020. Debunking handbook 2020.

Bing Liu. 2012. *Sentiment analysis and opinion mining*. Springer Nature.

Phillip May, Deutsche Telekom AG, and deepset GmbH. 2023. German BERT large paraphrase cosine. https://huggingface.co/deutsche-telekom/gbert-large-paraphrase-cosine. Accessed: 2024-05-13.

MDR. 2020. 2020: Die Chronik der Corona-Krise. https://www.mdr.de/nachrichten/jahresrueckblick/corona-chronik-chronologie-coronavirus-102.html. Accessed: 2024-04-12.

Nicholas Monath, Kumar Avinava Dubey, Guru Guruganesh, Manzil Zaheer, Amr Ahmed, Andrew McCallum, Gokhan Mergen, Marc Najork, Mert Terzihan, Bryon Tjanaka, Yuan Wang, and Yuchen Wu. 2021. Scalable hierarchical agglomerative clustering. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, page 1245–1255, New York, NY, USA. Association for Computing Machinery.

Nicholas Monath, Manzil Zaheer, and Andrew McCallum. 2023. Online level-wise hierarchical clustering. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 1733–1745, New York, NY, USA. Association for Computing Machinery.

NORD24. 2020. "Schnutenpulli" ist das plattdeutsche Wort des Jahres. `https://www.nord24.de/der-norden/schnutenpulli-ist-das-plattdeutsche-wort-des-jahres-44562.html`. Accessed: 2024-05-15.

Katrin Ortmann, Adam Roussel, and Stefanie Dipper. 2019. Evaluating off-the-shelf NLP tools for German. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019): Long Papers*, pages 212–222, Erlangen, Germany. German Society for Computational Linguistics & Language Technology.

Thomas Proisl. 2018. SoMeWeTa: A part-of-speech tagger for German social media and web texts. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 665–670, Miyazaki. European Language Resources Association ELRA.

Thomas Proisl and Peter Uhrig. 2016. SoMaJo: State-of-the-art tokenization for German web and social media texts. In *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task*, pages 57–62, Berlin. Association for Computational Linguistics.

Tina D Purnat, Paolo Vacca, Christine Czerniak, Sarah Ball, Stefano Burzo, Tim Zecchin, Amy Wright, Supriya Bezbaruah, Faizza Tanggol, Ève Dubé, Fabienne Labbé, Maude Dionne, Jaya Lamichhane, Avichal Mahajan, Sylvie Briand, and Tim Nguyen. 2021. Infodemic signal detection during the COVID-19 pandemic: Development of a methodology for identifying potential information voids in online conversations. *JMIR Infodemiology*, 1(1):e30971.

Juan Enrique Ramos. 2003. Using TF-IDF to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*.

Markus Reiter-Haas, Beate Klösch, Markus Hadler, and Elisabeth Lex. 2023. Polarization of opinions on COVID-19 measures: Integrating Twitter and survey data. *Social Science Computer Review*, 41(5):1811–1835.

Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 704–714, Seattle, Washington, USA. Association for Computational Linguistics.

Robert Koch-Institut. 2024. 7-Tage-Inzidenz der COVID-19-Fälle in Deutschland.

Francisco Rowe, Michael Mahony, Eduardo Graells-Garrido, Marzia Rango, and Niklas Sievers. 2021. Using Twitter to track immigration sentiment during early stages of the COVID-19 pandemic. *Data & Policy*, 3:e36.

Abraham C Sanders, Rachael C White, Lauren S Severson, Rufeng Ma, Richard McQueen, Haniel C Alcântara Paulo, Yucheng Zhang, John S Erickson, and Kristin P Bennett. 2021. Unmasking the conversation on masks: Natural language processing for topical sentiment analysis of COVID-19 Twitter discourse. *AMIA Jt Summits Transl Sci Proc*, 2021:555–564.

Arnaldo Santoro, Alessandro Galeazzi, Teresa Scantamburlo, Andrea Baronchelli, Walter Quattrociocchi, and Fabiana Zollo. 2023. Analyzing the changing landscape of the Covid-19 vaccine debate on Twitter. *Social Network Analysis and Mining*, 13(1):115.

H. Schmid. 1999. *Improvements in Part-of-Speech Tagging with an Application to German*, pages 13–25. Springer Netherlands, Dordrecht.

Thomas Schmidt, Jakob Fehle, Maximilian Weissenbacher, Jonathan Richter, Philipp Gottschalk, and Christian Wolff. 2022. Sentiment analysis on Twitter for the major German parties during the 2021 German federal election. In *Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)*, pages 74–87, Potsdam, Germany. KONVENS 2022 Organizers.

Skipper Seabold and Josef Perktold. 2010. Statsmodels: Econometric and statistical modeling with Python. In *9th Python in Science Conference*, pages 92–96.

Ben Shneiderman. 2003. The eyes have it: A task by data type taxonomy for information visualizations. In *The craft of information visualization*, pages 364–371. Elsevier.

Suzanna Sia, Ayush Dalmia, and Sabrina J. Mielke. 2020. Tired of topic models? clusters of pretrained word embeddings make for fast and good topics too! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1728–1736, Online. Association for Computational Linguistics.

Margaret C. Stewart and Christa L. Arnold. 2018. Defining social listening: Recognizing an emerging dimension of listening. *International Journal of Listening*, 32(2):85–100.

Julia Maria Struß, Melanie Siegel, Josef Ruppenhofer, Michael Wiegand, and Manfred Klenner. 2019. Overview of GermEval task 2, 2019 shared task on the identification of offensive language. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*, pages 352 – 363.

Tagesschau. 2020a. "Eine fehlgeleitete Diskussion". https://www.tagesschau.de/inland/corona-maskenpflicht-virologe-101.html. Accessed: 2024-04-30.

Tagesschau. 2020b. Maskenpflicht in allen Bundesländern. https://www.tagesschau.de/inland/corona-maskenpflicht-103.html. Accessed: 2024-04-30.

Alexei Vinokourov, Nello Cristianini, and John Shawe-Taylor. 2002. Inferring a semantic representation of text via cross-language correlation analysis. In *Advances in Neural Information Processing Systems*, volume 15. MIT Press.

Emily K. Vraga, Ullrich K. H. Ecker, Iris Žeželj, Aleksandra Lazić, and Arina A. Azlan. 2023. *To Debunk or Not to Debunk? Correcting (Mis)Information*, pages 85–98. Springer International Publishing, Cham.

We Are Social, DataReportal, and Meltwater. 2024a. Global social network penetration rate as of January 2024, by region. https://www.statista.com/statistics/269615/social-network-penetration-by-region/. Accessed: 2024-04-12.

We Are Social, DataReportal, and Meltwater. 2024b. Most popular reasons for internet users worldwide to use social media as of 3rd quarter 2023. https://www.statista.com/statistics/715449/social-media-usage-reasons-worldwide/. Accessed: 2024-04-30.

Silvan Wehrli, Bert Arnrich, and Christopher Irrgang. 2023. German text embedding clustering benchmark. In *Proceedings of the 19th Conference on Natural Language Processing (KONVENS 2023)*, pages 187–201, Ingolstadt, Germany. Association for Computational Lingustics.

Silvan Wehrli, Christopher Irrgang, Mark Scott, Bert Arnrich, and T. Sonia Boender. 2024. The role of the (in)accessibility of social media data for infodemic management: a public health perspective on the situation in the European Union in March 2024. *Frontiers in Public Health*, 12.

Becky K White, Arnault Gombert, Tim Nguyen, Brian Yau, Atsuyoshi Ishizumi, Laura Kirchner, Alicia León, Harry Wilson, Giovanna Jaramillo-Gutierrez, Jesus Cerquides, Marcelo D'Agostino, Cristiana Salvi, Ravi Shankar Sreenath, Kimberly Rambaud, Dalia Samhouri, Sylvie Briand, and Tina D Purnat. 2023. Using machine learning technology (early artificial intelligence–supported response with social listening platform) to enhance digital social understanding for the COVID-19 infodemic: Development and implementation study. *JMIR Infodemiology*, 3:e47317.

Elisabeth Wilhelm, Isabella Ballalai, Marie-Eve Belanger, Peter Benjamin, Catherine Bertrand-Ferrandis, Supriya Bezbaruah, Sylvie Briand, Ian Brooks, Richard Bruns, Lucie M Bucci, Neville Calleja, Howard Chiou, Abhinav Devaria, Lorena Dini, Hyjel D'Souza, Adam G Dunn, Johannes C Eichstaedt, Silvia M A A Evers, Nina Gobat, Mika Gissler, Ian Christian Gonzales, Anatoliy Gruzd, Sarah Hess, Atsuyoshi Ishizumi, Oommen John, Ashish Joshi, Benjamin Kaluza, Nagwa Khamis, Monika Kosinska, Shibani Kulkarni, Dimitra Lingri, Ramona Ludolph, Tim Mackey, Stefan Mandić-Rajčević, Filippo Menczer, Vijaybabu Mudaliar, Shruti Murthy, Syed Nazakat, Tim Nguyen, Jennifer Nilsen, Elena Pallari, Natalia Pasternak Taschner, Elena Petelos, Mitchell J Prinstein, Jon Roozenbeek, Anton Schneider, Varadharajan Srinivasan, Aleksandar Stevanović, Brigitte Strahwald, Shabbir Syed Abdul, Sandra Varaidzo Machiri, Sander van der Linden, Christopher Voegeli, Claire Wardle, Odette Wegwarth, Becky K White, Estelle Willie, Brian Yau, and Tina D Purnat. 2023. Measuring the burden of infodemics: Summary of the methods and results of the fifth WHO infodemic management conference. *JMIR Infodemiology*, 3:e44207.

Peter Willett. 1988. Recent trends in hierarchic document clustering: A critical review. *Information Processing & Management*, 24(5):577–597.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

X. 2023a. Academic Research access. https://web.archive.org/web/20230202074709/https://developer.twitter.com/en/products/twitter-api/academic-research. Accessed: 2024-04-29.

X. 2023b. Developer agreement and policy. https://developer.twitter.com/en/developer-terms/agreement-and-policy. Accessed: 2024-05-12.

X. 2024. Tweet counts. https://developer.twitter.com/en/docs/twitter-api/tweets/counts/introduction. Accessed: 2024-04-29.

Jiaming Xu, Peng Wang, Guanhua Tian, Bo Xu, Jun Zhao, Fangyuan Wang, and Hongwei Hao. 2015. Short text clustering via convolutional neural networks. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Process-*

ing, pages 62–69, Denver, Colorado. Association for Computational Linguistics.

Andrea Zielinski, Calvin Spolwind, Henning Kroll, and Anna Grimm. 2023. A dataset for explainable sentiment analysis in the German automotive industry. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 138–148, Toronto, Canada. Association for Computational Linguistics.

A Additional Classification Results of the Sentiment Classification

	neutral	147	22	6
GT	negative	20	149	3
	positive	12	6	30
		neutral	negative	positive
			P	

Table 3: Confusion matrix showing the test set results of the sentiment classification of face mask-related X posts for GBERT$_{\text{TAPT+SFT}}$. **GT** denotes the ground truth and **P** the model's predictions.

Class	**Precision**	**Recall**	**F1**
neutral	84.18%	86.63%	85.39%
negative	82.12%	84.00%	83.05%
positive	76.92%	62.50%	68.97%

Table 4: Per-class test set results of the sentiment classification of face mask-related X posts for GBERT$_{\text{TAPT+SFT}}$.

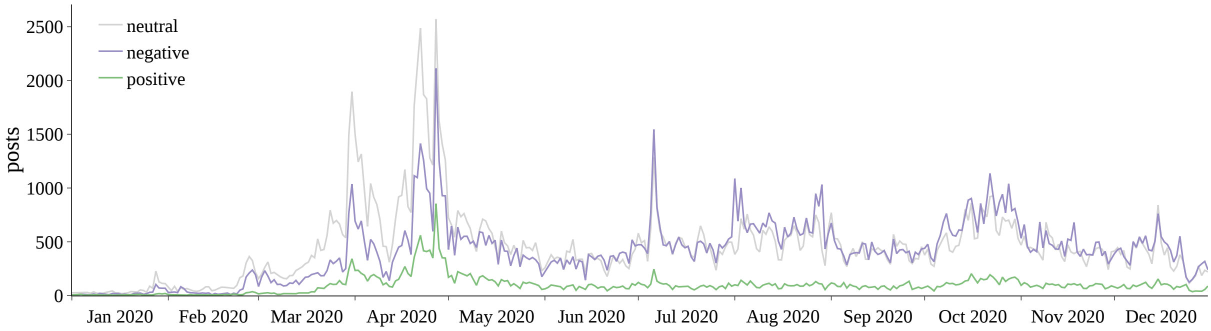

Figure 4: The number of German X posts with *neutral*, *negative*, and *posititve* sentiments on the topic of face masks per day in 2020. GBERT$_{\text{TAPT+SFT}}$ was used for sentiment classification.