Automatic Identification of COVID-19-related Narratives in German Telegram Channels and Chats

Philipp Heinrich[†] **Andreas Blombach**[†] **Bao Minh Doan Dang**[†]
Leonardo Zilio[†] **Linda Havenstein**[‡] **Nathan Dykes**[†]
Stephanie Evert[†] **Fabian Schäfer**[‡]
[†]Chair of Computational Corpus Linguistics [‡]Chair of Japanese Studies
Friedrich-Alexander-Universität Erlangen-Nürnberg
[†]Bismarckstr. 6, 91054 Erlangen [‡]Artilleriestr. 70, 91052 Erlangen
{firstname.lastname}@fau.de

Abstract

We are concerned with mapping the discursive landscape of conspiracy narratives surrounding the COVID-19 pandemic. In the present study, we analyse a corpus of more than 1,000 German Telegram posts manually tagged with 14 conspiracy and conspiracy-related narrative labels by three independent annotators. Since emerging narratives on social media are short-lived and notoriously hard to track, we experiment with different state-of-the-art approaches to few-shot and zero-shot text classification. We report performance in terms of ROC-AUC and in terms of optimal F_1, and compare fine-tuned methods with off-the-shelf approaches and human performance.

Keywords: COVID-19, User-generated Content, Zero-shot Text Classification, Few-shot Text Classification

1. Introduction and Related Work

In early 2020, shortly after declaring the spread of the new coronavirus SARS-CoV-2 a pandemic, the WHO also warned about an 'infodemic', a surge of disinformation, conspiracy narratives and misrepresentation of medical facts and political processes surrounding COVID-19. This came as no surprise, as "belief in conspiracy theories is stronger under conditions of uncertainty" and "when events are especially large-scale or significant" (Douglas et al., 2019); it thrives in times of crises and information vacuums. In addition, the 'connectedness' of the internet and especially social media have contributed to the spread of conspiracy narratives, as it turns the conspiracy narrative baseline of "'everything [being] connected' into reality" and the "interpretative logic of conspiracy theories [mirror] the ordering principle of the World Wide Web" (Butter, 2018).

Studies in Germany confirm a prevalence of conspiracy beliefs (Kuhn et al., 2021), with one in five citizens believing that the dangers of SARS-CoV-2 have been intentionally exaggerated to deceive the public (dimap, 2020) and the same proportion of people agreeing that 'Many numbers and statistics concerning COVID-19 are forged' (Institut für Demoskopie Allensbach, 2022). As a large part of the related conspiracy narrative and disinformation discussion has migrated to largely unmoderated platforms such as Telegram, it is difficult to oversee radicalisation processes and potentially hazardous developments within the scene and beyond.

Automatically identifying misinformation such as fake news, conspiracy narratives, or general *drivel*[1] is notoriously difficult. One of the major bottlenecks is the lack of suitable training (and evaluation) data, especially in a discursive landscape where narratives are evolving quickly. In the present study, we use a collection of 1099 posts scraped from openly accessible and popular COVID-19-themed Telegram channels and chat groups, which has been labelled manually by domain experts (see Section 2. To bypass the problem of sparse categories, we experiment with approaches that leverage label descriptions created by domain experts and with approaches to zero-shot and few-shot classification (i.e. techniques that use no examples or just a few examples of training data, see Section 3).

Related work in Natural Language Processing (NLP) often focuses on identifying fake news and "rumours" (Li and Zhou, 2020), which are related to, but different from, conspiracy theories. Moreover, the task is usually a yes/no classification ("drivel" vs. "no drivel") to assist moderation on social media (Moffitt et al., 2021). For deeper linguistic or computational social science analyses, or in order to apply counter-measures targeted to specific narratives, this binary approach is insufficient. Thus, our aim is to identify different groups of conspiracy-related or conspiracy-adjacent content. Previous work also attempted to automatically identify new conspiracy theories early on (Shahsavari et al., 2020; Marcellino et al., 2021).

With the advent of large language models (LLMs),

[1] German *Geschwurbel*, meaning conspiracy-related or conspiracy-adjacent content. We use *drivel* as a cover term for any such content in this paper.

LREC-COLING 2024, pages 1932–1943
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

especially BERT-like models (Devlin et al., 2019) and generative models such as the GPT series (Radford et al., 2018), researchers noticed that these LLMs contain a wealth of information about language and lexical semantic relations. This rich lexical information meant that textual relations can be predicted even without directly training the model for a task and gave rise to new approaches to zero-shot text classification.

An early benchmark for zero-shot text classification based on natural language inference (NLI) was proposed by Yin et al. (2019). By using entailment predictions between texts and hypotheses, the probability of the entailment can serve as a proxy for classification. Using this approach, Barker et al. (2021) achieved good results for zero-shot single-label classification of English texts. Similarly, large generative models can be used to predict text classifications based on a given prompt (Han et al., 2022).

In our study, we adopt an NLI-based approach using models for multi-label text classification based on DeBERTa (He et al., 2020) and RoBERTa (Liu et al., 2019). We also adopt a sentence-similarity-based approach (Reimers and Gurevych, 2019) to evaluate similarities between posts and label descriptions. Finally, we test the generative capacity of ChatGPT4 (OpenAI, 2023) in a zero-shot setting.

The main contributions of this study are the following:

- The collection of a large corpus of Telegram posts and the annotation of a sample with different COVID-19-related narratives, as presented in Section 2. The whole data set was used for training a GBert-based masked language model adapted to user-generated content related to the COVID-19 pandemic.
- A battery of text classification experiments on user-generated content, including classic machine-learning algorithms, zero-shot and few-shot classification (Section 3). We also provide a stratified split into train, development, and test sets that can be used for text classification.
- A comparison of the performance of machine learning algorithms with each other and with that of human annotators on the task of detecting COVID-19-related narratives (Section 4).

2. Corpus and Categorisation

In 2020 – as YouTube, Facebook, and others became more aggressive in cracking down on the spread of disinformation – sceptics, lockdown critics and conspiracy theorists found themselves in need of a new social media network. While new platforms and hosting services were set up for video streaming, a large part of the text- and image-based discussion migrated to the messaging and microblogging platform Telegram (Lamberty et al., 2022; Holnburger et al., 2022), widely known for its lack of moderation. Telegram channels and groups have thus become one of the most important data sources for studying conspiracy theories.

To build our corpus, we first scraped the channels of several well-known figures in the COVID-19 conspiracy scene using Telegram's own export function. Since channels often interact with each other (e.g. by forwarding messages), we proceeded to scrape frequently mentioned channels with large numbers of followers, thus iteratively increasing the scope and size of the corpus. This approach was supplemented by channel statistics available on the web.[2]

Our full corpus contains over 200 different Telegram channels (with follower counts ranging from a few thousand to over 300,000), as well as over 100 public group chats from January 2020 up to and including July 2022. These figures translate to a total of over 13 million posts, amounting to almost 400 million tokens. Upon request, interested researchers can be given access to search the corpus online.[3]

2.1. A masked language model of conspiratorial talk

With several hundreds of million tokens of running text, the corpus itself can be used to adapt a masked language model to the domain of conspiratorial talk on German Telegram. We use gbert-large[4] as a base model and fine-tuned it using the `transformers` library in Python. The model is available via Huggingface hub[5] and can be used for fine-tuning to a task such as text classification. Note that we do not use the model here, since we do not have a suitable data set for fine-tuning (fine-tuning a masked language model from scratch for text classification needs more examples than the couple of examples we provide with our annotation below).

2.2. Sample

In order to obtain a sample for the manual annotation of conspiracy narratives and related content, we first excluded forwarded messages and posts containing images, videos or polls (as we are only

[2] https://telemetr.io/en/channels?languages=de

[3] Please contact the first author to get access at https://corpora.linguistik.uni-erlangen.de/cqpweb/schwurpus_v2/.

[4] https://huggingface.co/deepset/gbert-large

[5] https://huggingface.co/ausgerechnet/schwurpert

concerned with content in textual form). Furthermore, we required a minimum length from the posts ($\geq$ 400 characters, excluding URLs). We then drew a sample from the filtered corpus, stratified by month, channel/group and number of messages, resulting in 1099 posts by 343 individual users in 143 different channels/groups from January 2020 to March 2022. This set of posts contains an average of 180 tokens distributed across 11.4 sentences.

2.3. Narratives

In order to annotate relevant narratives, it was necessary to develop a categorisation scheme. Our scheme is based on previous research (Institut für Demoskopie Allensbach, 2022; Kuhn et al., 2021), domain knowledge and close reading of excerpts from our corpus prior to sampling and was further refined during the annotation process. The categorisation scheme is hierarchical and contains a total of 18 narrative groups subdivided into 63 fine-grained narratives. It includes descriptions and examples for each narrative and is available online.[6] Description sentences were derived from the annotation guidelines by domain experts and are meant to represent concise summaries of the narratives. We include narratives specific to COVID-19, such as 'COVID-19 is no more dangerous than the common flu' or 'The pandemic serves to implement the Great Reset'[7] as well as previously existing narratives such as 'New World Order' or 'sheeple'.[8]

Since many fine-grained narratives are very infrequent in the sample (and a few are not present at all), we only use the (slightly adapted) narrative groups for the classification task in this papers. To give an idea of the narrative contents, brief descriptions in English are provided below; see also Table 1 for an overview.

Pseudo-pandemic : narratives that downplay the danger of COVID-19, deny its existence or feed doubts about the official narrative of the pandemic.

Criticism of countermeasures : narratives claiming that pandemic response efforts are illegal, more dangerous than the virus (e.g. masks causing illness) or that they discriminate against sceptics and people who refuse to wear masks, be tested or vaccinated.

Alternative treatments : narratives about repurposed drugs or other 'miracle cures' against COVID-19 that are allegedly withheld from the population.

narrative	sentences
pseudo-pandemic	13
criticism of countermeasures	12
alternative treatments	6
vaccine hazards	16
COVID-19 conspiracies	26
other conspiracies	9
QAnon	6
group-focused enmity	19
sheeple	2
millenarianism	4
state as an enemy	5
indoctrination	7
esotericism & pseudo-science	8
other drivel	9
"no drivel"	0

Table 1: List of narrative groups with number of description sentences.

Vaccine hazards : narratives portraying COVID-19 vaccines as insufficiently tested, unsafe, or even dangerous.

COVID-19 conspiracies : conspiracy theories claiming that some hidden agenda is behind the pandemic, e.g. Bill Gates aiming to reduce the world population or to inject people with microchips to control them, or the pandemic serving to destroy the economy, to achieve climate change goals or to produce profit for big corporations and powerful elites.

Other conspiracies : pre-existing conspiracy theories not specific to COVID-19 – chemtrails, claims about false flag operations, mind control, all-powerful secret societies etc.

QAnon : narratives about an anonymous individual called *Q* and his claims of insider knowledge about highly classified U.S. government documents, a Satanic cabal operating a global child sex trafficking ring, and Donald Trump's secret fight against this cabal.

Group-focused enmity : all forms of racism, xenophobia, Islamophobia, homophobia, misogyny etc., as well as the far-right conspiracy theories *Great Replacement* (claiming a plot to replace the ethnic white population with non-white immigrants, especially Muslims) and *BRD GmbH* (claiming that Germany never ceased to be controlled by the Allies after World War II, and rejecting the constitution and legitimacy of the modern German state in favor of the German Reich).

[6] https://github.com/fau-klue/infodemic

[7] The attainment of political or economic world domination by global financial elites, represented by the World Economic Forum and Klaus Schwab.

[8] We use the term 'narrative' very loosely throughout this paper. Some of the fine-grained categories are perhaps better thought of as building blocks of narratives.

Sheeple : covers the single narrative (often accompanying conspiracy narratives) that most people have no idea what is really going on, because they are brainwashed or choose to live in ignorance.

Millenarianism : narratives about an upcoming day or time of great change or reckoning when the group's beliefs will be validated and/or its enemies will be defeated.

State as an enemy : narratives questioning the status of democracy – by assuming a deep state that holds the real power, by accusing free actors of being covert agents of the system, by doubting election results, or by accusing the state of having dictatorial features.

Indoctrination : narratives claiming that the (mainstream) media is controlled by the state or a powerful group and/or that it lies, censors information and indoctrinates the population, as well as narratives about cancel culture and the death of free speech.

Esotericism & pseudo-science : pseudo-scientific claims about medicine (e.g. disbelief in the existence of viruses in general) or alternative medicine, as well as various esoteric practices and beliefs (healing crystals, mediums, auras etc.).

Other drivel : includes narratives with very low prevalence: climate change denial and narratives about a man-made origin of COVID-19.[9]

2.4. Manual annotation

Three of the authors individually annotated the full sample on two levels: whether a post contains *drivel* (conspiracy-related or conspiracy-adjacent content), and if so, which specific narratives it contains. Note that multiple categories can be assigned to the same post. In a subsequent adjudication process, we resolved all disagreements to arrive at a final gold standard.

Figure 1 shows the prevalence of each narrative group in the sample, as well as the number of posts containing no (or no clearly identifiable) *drivel* (52.5%). If several narratives belonging to the same narrative group occur in the same post, they are counted as only one instance of the group.

	Rater 1	Rater 2	Rater 3
Gold	0.84	0.78	0.67
Rater 1		0.67	0.57
Rater 2			0.60

Table 2: Pairwise Cohen's κ for the initial classification as *drivel* or not (Fleiss' κ excluding the gold standard: 0.61).

Table 2 shows pairwise inter-annotator agreement for the first level of annotation. As evident from the moderate to good values, even this binary classification is often difficult for domain experts. Fleiss κ for individual narrative groups ranges from .32 to .83, with a mean of .59 and a standard deviation of .13.

3. Automatic Classification of Posts

Automatic classification of posts is operationalised as a multi-label document classification problem: the task is to identify which narratives, if any, are mentioned in a post. Since the annotation of training data is very resource-intensive and our categorisation scheme contains a large number of narratives, only a handful of positive examples can be used for training ("few-shot classification", see Section 3.1). Alternatively, we can use zero-shot classification (i.e. classification using no training data, Section 3.2), which derives its predictions from the text to be classified and the semantics of the category label (in our case: the description of a narrative). Note that the descriptions can also be used in supervised approaches by including them as positive examples in the training data.

For evaluation, we treat the problem as separate binary document classification task for each narrative group, so we can quantify performance per group by area under the receiver operating characteristic curve (ROC-AUC). This is a reasonable choice since the classification threshold determining sensitivity (i. e. recall) and specificity (or, alternatively, precision) can be set in a task-specific way, e.g. opting for high recall when using the classifier as a filter whose results are checked manually afterwards.

The complete annotated corpus is split in a stratified fashion into training, development and test sets, with a ratio of $60 : 15 : 25$.[10] All models that require training are trained on the training split. Suitable cut-off values for optimal F_1 are found on the development set. Measures provided in Table 3 (and Table 4 in the appendix) are derived from the test set.

[9] Note that the latter narratives were much less important in German-language discourse than, for example, in the U.S. While many people represented in our corpus might agree with the statement "COVID-19 was created in a laboratory", the question of *how* the virus is used by powerful people or organisations is usually more pertinent.

[10] See https://github.com/fau-klue/infodemic for the complete corpus and the split.

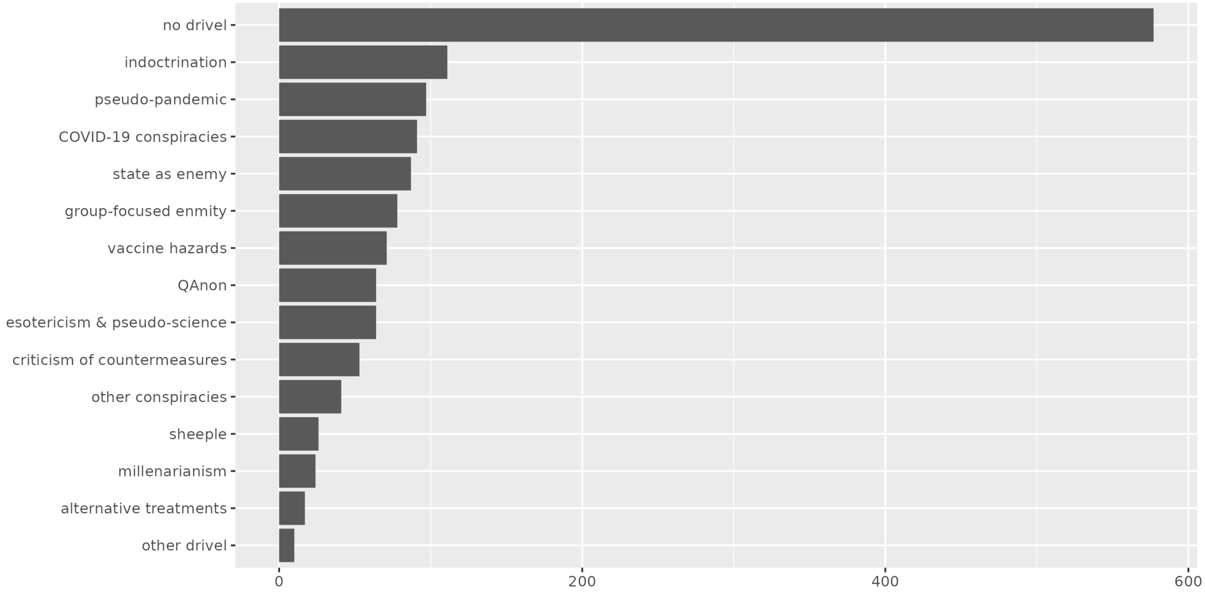

Figure 1: Distribution of labels of narrative groups in the annotated sample.

3.1. Supervised prediction

ML baseline In order to have proper baselines for our few-shot and zero-shot approaches, we first perform several multi-label experiments with standard machine learning classifiers. We use logistic regression (LR) and a support vector machines (SVM) with a tf.idf weighted unigram bag-of-words feature matrix and perform experiments using `scikit-learn` (Pedregosa et al., 2011).

An additional question relating to supervised machine learning classification is whether adding a small number of description texts into training data can improve model performance (allowing a bag-of-words model to directly learn keywords from the descriptions). Therefore, we conduct our experiments as follows: in one round we use the training set (*posts only*), and in the other we extend training data with description sentences (*posts+descriptions*) to train our classification models.

Few-shot classification Since it is difficult and time-consuming to manually annotate narratives, and our data set is therefore comparatively small, few-shot learning is an obvious approach to generalise from only a small amount of training data for each label.

Tunstall et al. (2022) proposed SetFit, a framework for few-shot fine-tuning Sentence Transformers (Reimers and Gurevych, 2019). A pre-trained Sentence Transformers model is first fine-tuned on a number of contrastive pairs of labelled texts. This model is then used to encode the training data. Finally, a text classification head is trained using the encoded data. While state-of-the-art methods such as T-Few (Liu et al., 2022) may attain even better few-shot results, SetFit is competitive and has the advantages of not requiring prompts and being easier to train.

In our experiments, we fine-tuned different Sentence Transformers models, but quickly found paraphrase-multilingual-mpnet-base-v2[11] to be the best available base model. We also found that using the description sentences for our target labels as additional training data significantly improved model performance. We therefore report only these results for the SetFit approach. We include both “out of the box” results and results after (time-consuming) hyperparameter optimisation.

3.2. Zero-shot prediction

For zero-shot classification, we split the posts and narrative descriptions into sentences.[12] For a post p and a narrative n, we can calculate one score $s\,(s_i, s_j)$ for each sentence pair

$$(s_i, s_j) \text{ with } s_i \in S_p, s_j \in S_n,$$

where S_p comprises the sentences of post p and S_n the sentences of narrative n; see below for the exact procedures to get scores.

The individual sentences of a narrative description usually represent different ways in which the narrative can be expressed. Since we are primarily interested in whether a given conspiracy narrative is present in the post or not (rather than how it is

[11] https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2

[12] Due to API limitations, ChatGPT4 experiments were conducted in a different way.

formulated or whether the post is exclusively about this narrative), it seems reasonable to take the maximum over the scores of all sentence pairs as the overall score for post p and narrative n:[13]

$$\text{score}\,(p, n) = \max_{s_i \in S_p, s_j \in S_n} \left(s(s_i, s_j)\right)$$

Note that we could also opt for comparing entire posts to each sentence of the narrative descriptions or all sentences in a post to the whole narrative descriptions. However, these strategies were consistently outperformed by the approach above, and we exclude those results for the sake of brevity.

Sentence-similarity zero-shot A cheap approach to zero-shot classification can be constructed by looking at similarities between posts and narratives at the sentence level. We encode all sentences using a multi-lingual SBERT model (Reimers and Gurevych, 2019). Here, we use the Python `SentenceTransformer` library[14] and opt for distiluse-base-multilingual-cased-v1 (Reimers and Gurevych, 2020), which yielded good results for German CMC in a pre-study, even compared to specialised German embeddings. We then use the cosine similarity between the SBERT sentence embeddings $\mathbf{e}_{s_i}$ and $\mathbf{e}_{s_j}$ as a score:

$$s(s_i, s_j) = \cos\left(\mathbf{e}_{s_i}, \mathbf{e}_{s_j}\right) \quad \forall s_i \in S_p, s_j \in S_n$$

NLI zero-shot Yin et al. (2019) pioneered the use of natural language inference (NLI) models for zero-shot text classification. The main idea is that if a model can predict whether a hypothesis is semantically entailed from a text, it can also be used for text classification with previously unseen labels. In this study, instead of just using single- or few-word labels, we can make use of a detailed textual description of the label, consisting of several sentences. We tested entailment hypotheses for each sentence from the description against each sentence from a given post. The following hypothesis template was used: “In diesem Satz geht es um das Thema {}.” [This sentence is about {}.], where “{}” was replaced with sentences from label descriptions.

We used four different models pre-trained for NLI, all of which are available on Huggingface[15] and are either specifically trained for the German language or, in the case of multilingual models, include German in their training data:

- gbert-large-nli[16]: This is the only model specifically fine-tuned for German. It uses the 10KGNAD data set (Schabus et al., 2017) on top of the German BERT-large model (Chan et al., 2020).

- xlm-roberta-large-xnli[17]: This model was fine-tuned on xlm-roberta-large (Conneau and Lample, 2019) using the XNLI Corpus to perform NLI for 15 languages (Conneau et al., 2018).

- mDeBERTa-v3-base-xnli[18] (Laurer et al., 2023): This model was fine-tuned for NLI on top of mDeBERTa-base v3(He et al., 2022), also using the XNLI Corpus as basis.

- mDeBERTa-v3-base-xnli-2mil7[19] (Laurer et al., 2023): This model also used mDeBERTa-base v3, but fine-tuned it for NLI using the XNLI Corpus translated to 26 languages, containing a total of 2.7 million text pairs.

ChatGPT4 To obtain predictions for our test set from ChatGPT, we used OpenAI’s (paid) API via Python.[20] The model (gpt-4-0613) first received a system prompt (in English) to steer its classification behaviour. This included information about the task, the expected input and output format, as well as the same label description sentences (in German) used in our previous experiments.

> You will be provided with German Telegram posts from people who are potentially spreading COVID-19 misinformation and conspiracy theories.
>
> Posts will be delimited with ∼∼∼∼ characters. Each post will be preceded by a unique numeric id on the first line.
>
> Your job is a multi-label classification task. Each post can belong to one or more of 14 different classes. If none of these classes apply, a post is to be labelled "kein_Geschwurbel".
>
> For each class or label, there are multiple description sentences in German in the

[13] In a different setting, other aggregation procedures could also be reasonable, such as taking the overall mean or the mean of the maximum scores.

[14] https://www.sbert.net/

[15] https://huggingface.co/

[16] https://huggingface.co/svalabs/gbert-large-zeroshot-nli. A more detailed description of this model is available in German at: https://focus.sva.de/big-data-analytics/zeroshot-klassifikation/

[17] https://huggingface.co/joeddav/xlm-roberta-large-xnli

[18] https://huggingface.co/MoritzLaurer/mDeBERTa-v3-base-mnli-xnli

[19] https://huggingface.co/MoritzLaurer/mDeBERTa-v3-base-xnli-multilingual-nli-2mil7

[20] https://github.com/openai/openai-python

> form "label: description". These label descriptions, one per line, try to encapsulate the essence of several subclasses.
>
> The labels and their descriptions are:
>
> Impfung_ist_gefährlich: Für die Herstellung neuartiger mRNA-Impfstoffe gegen das Coronavirus (wie von BioNTech oder Moderna) werden menschliche Embryonen oder abgetriebene Föten benutzt.
> Impfung_ist_gefährlich: Die Impfung gegen COVID-19 ist nicht ausreichend getestet worden und deshalb nicht sicher. Mögliche Langzeitschäden durch die Impfung lassen sich nicht ausschließen.
>
> ...
>
> Millenarismus: Am Tag X wird das Volk sich erheben, die Regierung stürzen und zur Rechenschaft ziehen.
> Schlafschafe: Ein großer Teil der Bevölkerung ist gehirngewaschen und hat keine Ahnung, was wirklich in der Welt vor sich geht.
> Schlafschafe: Schlafschafe lassen sich von den Medien blenden und erdulden alles wie brave Schafe, anstatt sich zu wehren.
>
> Classify each post according to the labels and descriptions above. Use only these labels. Multiple labels per post are possible.
>
> Provide your output in json format with ids and all applicable labels.

Posts were submitted in batches as user prompts. Since the chosen model could only process 8,192 tokens at a time and the system prompt – which had to precede every batch of posts – required over 5,500 tokens (due to the amount and length of the description sentences), we could only submit a few posts at a time. Paired with OpenAI's rate limit of 10,000 tokens per minute, API outages and inconsistent output from the model, this process was not nearly as smooth as initially expected.[21]

4. Results

All our models yield separate scores for each narrative. Figure 2 shows the receiver-operating characteristic curves of the most prevalent narrative (*Indoctrination*) for all models. Average ROC-AUC

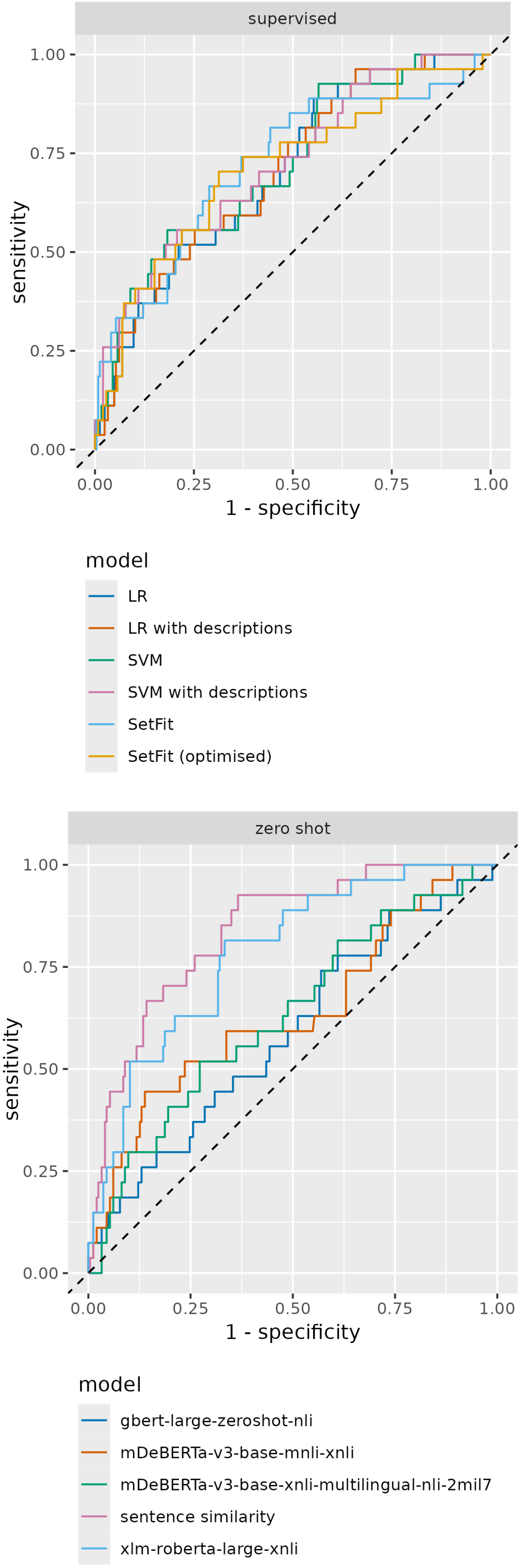

Figure 2: Receiver-operating characteristic curves for the most prevalent narrative *Indoctrination*, with supervised approaches on the top and zero-shot approaches on the bottom.

[21] In total, we spent approx. €10 for the actual experiments, but another €20 to €30 for unsuccessful trials. With current pricing, especially for models with a larger context window, this would be considerably cheaper. However, at least in our tests, a larger context window made the output even more inconsistent.

	model	avg.	avg. F_1	
		ROC-AUC	micro	macro
	Carina (student annotator)		0.76	0.72
	Laura (student annotator)		0.71	0.66
	ChatGPT4		0.52	0.43
supervised	SetFit (optimised)	0.79	**0.51**	0.37
	SetFit	0.76	0.50	0.36
	SVM	0.78	0.48	0.33
	SVM with descriptions	0.81	0.47	**0.38**
	LR with descriptions	0.80	0.45	0.35
	LR	0.77	0.40	0.28
zero shot	sentence similarity	**0.86**	0.40	0.32
	xlm-roberta-large-xnli	0.82	0.39	0.31
	mDeBERTa-v3-base-xnli-2mil7	0.69	0.22	0.21
	mDeBERTa-v3-base-xnli	0.70	0.20	0.21
	gbert-large-nli	0.58	0.15	0.15

Table 3: Macro-average ROC-AUC values for different models (first result column): We report macro averages of all 14 narratives exlcuding the negative category “no drivel”. The cheap approach using sentence similarity outperforms all other models. xlm-roberta-large-xnli yields by far the best results among the NLI zero-shot models. All other models are outperformed by our baselines (in particular SVM leverarging descriptions). Average optimal F_1 scores are reported on the right two result columns; the whole table is sorted by micro-average F_1. We include student annotators and ChatGPT4 (top), who outperform our models in terms of optimal F_1; here, only the few-shot approach (SetFit) beats our baselines.

across all narratives are reported in Table 3. Note that the supervised systems are trained with and can thus predict a label “no drivel”, which the zero-shot classifiers (except ChatGPT4) cannot. For reasons of comparability, we report the macro-average excluding the negative class; results for all labels are almost the same where applicable (see Table 4 for the complete picture).

In order to compare our systems in terms of precision and recall as well (or F_1, the harmonic mean of precision and recall), we have to determine a cut-off value for binary prediction of each narrative. We thus use the development set to choose a threshold that maximises F_1. In this scenario, we can also assign the label “no drivel” for all zero-shot classifiers: a post is classified as “no drivel” by a model if and only if no other label has been given to this post by the model. Table 3 (right) shows micro- and macro-average F_1-scores for all models across all 15 labels (using the optimal thresholds). We include student annotators and ChatGPT4 in this list and sort by descending micro average.

Table 3 shows that zero-shot approaches are very good at the detection of narratives: the sentence similarity approach beats all other systems in terms of average ROC-AUC and is clearly the best-performing model for predicting e.g. label ‘Indoctrination’, cf. Figure 2. Supervised ML approaches (including the few-shot classifier) clearly optimise trade-off between precision and recall and reach highest ranks when compared at optimal F_1; they can take the actual prevalence of each narrative into account, which zero-shot models cannot. Note that ChatGPT4 outperforms our models, likely because it is a much larger pre-trained model. Also note that all models are still far from human performance; this difference shows that this sort of prediction is a challenging task, and quantifies how much room for improvement is still left. We invite the research community to engineer better systems for solving the task on our data set.

5. Discussion

The present study is concerned with the identification of conspiracy narratives found on German social media (more specifically, Telegram). We have shown that leveraging class descriptions yields competitive results to state-of-the-art supervised learning techniques: the cheap sentence similarity approach outperforms all other approaches in terms of average ROC-AUC. Since the creation of training data labelled by domain experts is an expensive bottleneck of modern text classification – especially in times of fast-changing discursive landscapes –, zero-shot techniques represent an inexpensive and promising angle.

There are some obvious limits and caveats to our approach, whose investigation we leave for future research: Firstly, we did not systematically analyse the influence of description sentences on our classification procedure. Initial experiments indicate that a moderate amount of very specific sentences yields the best results. Similarly, we did not systematically experiment with different SBERT models in the sentence-similarity and few-shot approaches, and only included a handful of models in the NLI zero-shot approach. At this point, it seems that the general multi-language, multi-purpose models yield good results for our procedure, but fine-tuned embeddings might be a promising step. By providing a masked language model adapted to the whole corpus, we lay the foundation for further experiments. Last but not least, the ML classifiers improve with increasing numbers of training examples. A closer look at learning curves would thus be necessary in order to determine the point where supervised techniques start outperforming zero-shot classification techniques.

Lastly, our approach lends itself to a simple extension: descriptions of narratives can be very abstract on the one side ("arm-chair" descriptions) or very concrete on the other (actual surface realisations sampled from the corpus). From a practical point of view, it thus seems reasonable to start with basic descriptions of categories (such as the ones we used here) and extend the descriptions with sentences found in the classification procedure. Since no cut-off for inclusion can be determined *a priori*, this process should ideally be supervised (i.e., experts can select additional description sentences from n-best lists generated by the classifier).

6. Acknowledgements

This work has partially been funded by Volkswagen Foundation (project no. 98 776) and by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 466328567.

7. Supplementary Materials

7.1. Results per narrative group

We report complete results per narrative group in Table 4.

7.2. Ethical considerations

We collected posts from public Telegram channels which can be accessed by anyone with access to the world wide web, even without an account or a subscription. Users can thus not expect anonymity. Users generally do not use their real names in Telegram groups (for public figures running their own channel, like *Boris Reitschuster* or *Eva Herman*, this is obviously different). In a few cases, however, users did post their real identities and/or phone numbers (in terms of so-called "v-cards"). We stripped our dataset of these obvious identifiers before processing them further.

The manually annotated dataset, which we publish online, only contains channel and chat group names, no individual user names. The full corpus, on the other hand, is only available to other researchers upon request. Sharing the raw data and derived models among the research community is possible under German law (§60d UrhG).

Note that a working system for automatic narrative classification could be used for filtering or steering discussions (but given the current performance, such a system would sensibly only flag suspicious posts for human moderators). However, compared to the impact of current *generative* LLMs such as ChatGPT, practical relevance for malicious applications seems vanishingly low in practice. Ideally, such a model would be used by official moderators (of e. g. forums or comment sections).

Lastly, the present contribution contains academic experiments. For a productive deployment of any of the systems presented here, one would have to completely anonymise the training data beforehand.

				supervised						zero shot				
model →	Carina (student annotator)	Laura (student annotator)	ChatGPT4	SetFit (optimised)	SetFit	SVM	SVM with descriptions	LR with descriptions	LR	sentence similarity	xlm-roberta-large-xnli	mDeBERTa-v3-base-xnli-multilingual-nli-2mil7	mDeBERTa-v3-base-mnli-xnli	gbert-large-zeroshot-nli
narrative group ↓	ROC-AUC													
pseudo-pandemic				.74	.67	**.86**	.84	.83	.84	**.82**	.74	.72	.68	.65
crit. of countermeasures				.53	.58	.72	.71	**.73**	.72	**.87**	.84	.81	.74	.71
alternative treatments				.94	.79	.95	**.96**	**.96**	**.96**	.94	**.98**	.79	.77	.27
vaccine hazards				.84	.68	.85	**.89**	.88	.84	**.90**	.77	.80	.77	.66
COVID-19 conspiracies				.78	.77	.82	**.83**	**.83**	.80	**.85**	.74	.72	.77	.64
other conspiracies				.81	**.82**	.62	.66	.68	.63	.83	**.86**	.60	.59	.53
QAnon				**.81**	.78	.77	.77	.71	.71	**.84**	.66	.54	.65	.47
group-focused enmity				.80	.81	.81	**.89**	.88	.79	**.85**	.67	.73	.65	.47
sheeple				.77	**.79**	.76	.78	.77	.76	**.97**	.95	.82	.74	.53
millenarianism				.75	**.79**	.71	.75	.74	.68	.75	**.91**	.65	.83	.64
state as an enemy				.75	.75	.74	**.76**	**.76**	.75	.80	**.88**	.50	.50	.56
indoctrination				.71	**.72**	**.72**	**.72**	.70	.70	**.84**	.78	.63	.64	.59
esot. & pseudo-science				.84	.81	.87	**.89**	**.89**	.86	**.76**	.75	.63	.72	.55
other drivel				**1**	.80	.78	.94	.89	.75	**.99**	.96	.78	.77	.89
"no drivel"				.71	.70	.71	.70	.69	**.72**					
macro average				.79	.76	.78	**.81**	.80	.77	**.86**	.82	.69	.70	.58
– incl. "no drivel"				.78	.75	.78	**.81**	.80	.77					
narrative group ↓	(optimal) F_1													
pseudo-pandemic	.68	.58	.60	.27	.34	.45	.43	**.49**	.39	**.35**	.24	.20	**.35**	.23
crit. of countermeasures	.57	.50	.34	.22	.12	.15	.21	**.29**	.25	.20	**.27**	.22	.00	.13
alternative treatments	1	.89	.50	.44	**.67**	.00	.50	.29	.00	.13	**.30**	.11	.22	.02
vaccine hazards	.67	.74	.69	**.48**	.29	.46	.41	.24	.16	**.50**	.38	.38	.09	.20
COVID-19 conspiracies	.59	.61	.40	.38	.28	.37	**.41**	.40	.23	.30	**.34**	.25	.24	.17
other conspiracies	.56	.78	.27	**.18**	**.18**	.09	.09	.13	.08	**.29**	.21	.07	.05	.05
QAnon	.69	.74	.64	**.42**	.38	.24	.26	.31	.30	**.31**	.11	.11	.17	.12
group-focused enmity	.69	.74	.65	.22	**.55**	.44	.33	.33	.24	**.38**	.20	.20	.00	.00
sheeple	.77	.57	.20	.25	.29	**.46**	.35	.40	.43	.25	**.32**	.21	.24	.06
millenarianism	.40	.29	.00	.18	**.25**	.13	.22	.09	.08	.16	**.31**	.18	.44	.06
state as an enemy	.74	.71	.43	**.39**	.29	.23	.31	.27	.27	.19	**.47**	.13	.11	.17
indoctrination	.81	.70	.39	**.33**	.29	.32	.30	.22	.22	.12	**.34**	.14	.29	.21
esot. & pseudo-science	.83	.61	.57	.36	.36	.39	**.49**	.41	.37	.18	**.26**	.15	.15	.08
other drivel	1	.67	.10	**.67**	.40	.50	**.67**	**.67**	.50	**.67**	.07	.08	.03	.06
"no drivel"	.85	.81	.67	**.72**	.70	.71	.71	.71	**.72**	.79	**.84**	.78	.81	.71
micro average	.70	.65	.44	**.34**	.33	.32	.29	.28	.23	**.28**	**.28**	.17	.16	.13
– incl. "no drivel"	.76	.71	.52	**.51**	.50	.48	.47	.45	.40	**.40**	.39	.22	.20	.15
macro average	.71	.65	.41	.34	.33	.30	**.36**	.32	.25	**.29**	.27	.17	.17	.11
– incl. "no drivel"	.72	.66	.43	.37	.36	.33	**.38**	.35	.28	**.32**	.31	.21	.21	.15

Table 4: Complete results: ROC-AUC and (optimal) F_1 scores for each narrative group. We report micro and macro averages both for excluding and including prediction of "no drivel" (where applicable). Bold numbers represent best-performing systems for each type of model in terms of the respective score.

8. Bibliographical References

Ken Barker, Parul Awasthy, Jian Ni, and Radu Florian. 2021. IBM MNLP IE at CASE 2021 task 2: NLI reranking for zero-shot text classification. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 193–202.

Michael Butter. 2018. *»Nichts ist, wie es scheint«: Über Verschwörungstheorien*. Suhrkamp Verlag.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Infratest dimap. 2020. Corona und die Medien - Eine Studie im Auftrag des NDR – Magazin zapp - Mai 2020.

Karen M. Douglas, Joseph E. Uscinski, Robbie M. Sutton, Aleksandra Cichocka, Turkay Nefes, Chee Siang Ang, and Farzin Deravi. 2019. Understanding conspiracy theories. *Political Psychology*, 40(S1):7.

Xu Han, Weilin Zhao, Ning Ding, Zhiyuan Liu, and Maosong Sun. 2022. PTR: Prompt tuning with rules for text classification. *AI Open*, 3:182–192.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations*.

Josef Holnburger, Maheba Goedeke Tort, and Pia Lamberty. 2022. Q vadis? Zur Verbreitung von QAnon im deutschsprachigen Raum.

Institut für Demoskopie Allensbach. 2022. Politischer Radikalismus und die Neigung zu Verschwörungstheorien.

Sarah Anne Kezia Kuhn, Roselind Lieb, Daniel Freeman, Christina Andreou, and Thea Zander-Schellenberg. 2021. Coronavirus conspiracy beliefs in the German-speaking general population: endorsement rates and links to reasoning biases and paranoia. *Psychological Medicine*, pages 1–15.

Pia Lamberty, Josef Holnburger, and Maheba Goedeke Tort. 2022. Das Protestpotential während der COVID-19-Pandemie.

Qifei Li and Wangchunshu Zhou. 2020. Connecting the dots between fact verification and fake news detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1820–1825, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *arXiv preprint arXiv:2205.05638*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

William Marcellino, Todd C. Helmus, Joshua Kerrigan, Hilary Reininger, Rouslan I. Karimov, and Rebecca Ann Lawrence. 2021. Detecting conspiracy theories on social media: Improving machine learning to detect and understand online conspiracy theories. Technical report, RAND Corporation.

J. D. Moffitt, Catherine King, and Kathleen M. Carley. 2021. Hunting conspiracy theories during the COVID-19 pandemic. *Social Media + Society*, 7(3):1–17.

OpenAI. 2023. Gpt-4 technical report.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Shadi Shahsavari, Pavan Holur, Tianyi Wang, Timothy R. Tangherlini, and Vwani Roychowdhury. 2020. Conspiracy in the time of corona: automatic detection of emerging COVID-19 conspiracy theories in social media and the news. *Journal of Computational Social Science*, 3:279–317.

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. Efficient few-shot learning without prompts. *arXiv preprint arXiv:2209.11055*.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*, pages 3914–3923. Association for Computational Linguistics.

9. Language Resource References

Chan, Branden and Schweter, Stefan and Möller, Timo. 2020. *German's Next Language Model*. International Committee on Computational Linguistics.

Conneau, Alexis and Lample, Guillaume. 2019. *Cross-lingual language model pretraining*.

Conneau, Alexis and Rinott, Ruty and Lample, Guillaume and Williams, Adina and Bowman, Samuel R. and Schwenk, Holger and Stoyanov, Veselin. 2018. *XNLI: Evaluating Cross-lingual Sentence Representations*. Association for Computational Linguistics.

He, Pengcheng and Gao, Jianfeng and Chen, Weizhu. 2022. *DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing*.

Laurer, Moritz and van Atteveldt, Wouter and Casas, Andreu and Welbers, Kasper. 2023. *Less Annotating, More Classifying: Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT-NLI*. Cambridge University Press.

Reimers, Nils and Gurevych, Iryna. 2020. *Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation*. Association for Computational Linguistics.

Dietmar Schabus and Marcin Skowron and Martin Trapp. 2017. *One Million Posts: A Data Set of German Online Discussions*.