Dimensions of Drivel in German Telegram Posts
Manual Annotation and Predictive Power

Andreas Blombach[1], **Stephanie Evert**[1], **Linda Havenstein**[2], **Philipp Heinrich**[1]
[1]Lehrstuhl für Korpus- und Computerlinguistik [2]Lehrstuhl für Japanologie
Friedrich-Alexander-Universität Erlangen-Nürnberg
[1]Bismarckstr. 6, 91054 Erlangen [2]Artilleriestr. 70, 91052 Erlangen
{first.name}.{last.name}@fau.de

Abstract

Social media platforms abound with unsubstantiated claims and conspiratorial or conspiracy-adjacent discourse. In this paper, we aim to quantify the overall drivel-like quality of posts along six dimensions, each assessing a different aspect. Six student assistants annotated a random sample of 1,000 Telegram posts from well-known German conspiracy theorists, based on carefully formulated guidelines to ensure consistency. Each dimension was rated on a scale from 1 to 5, and inter-annotator agreement was evaluated using Krippendorff's α for ordinal scales, revealing moderate to substantial agreement across dimensions. We also report experiments on predicting the overall drivel-like quality of posts from these dimensions using a simple linear regression model. It shows that posts are considered drivel overall in particular when their contents appear distant from reality and when authors strongly assert their views.

Keywords: Telegram, conspiracy theories, drivel, annotation

1. Introduction

Conspiratorial, conspiracy-adjacent, esoteric, or pseudo-scientific discourse – often referred to in German as *Geschwurbel* – presents a persistent challenge in contemporary digital spaces, particularly on social media (Douglas et al., 2019). Such discourse is not always explicit or fully formed; rather, it frequently consists of vague insinuations, rhetorical devices, and partial references that can easily evade traditional classification systems.

Recent research has largely focused on identifying and categorising conspiracy narratives (Heinrich et al., 2024; Piskorski et al., 2025). However, with the exception of a few recent schemes (Piskorski et al., 2023), classification frameworks tend to ignore subtle cues such as rhetorical strategies, emotive language, or ambiguous references, which contribute to the conspiratorial tone without adhering to a concrete topic or narrative. Instead, much of this work has focused explicitly on narrative detection and binary classification of drivel. Many posts lacking a clear narrative – containing only hints or allusions – often appear vaguely conspiratorial but ultimately slip through the classification framework. Furthermore, since (some) narratives evolve over time (some fade while new ones emerge), relying solely on narrative structures may miss important cross-narrative patterns.

These observations motivated the development of a multi-stage scale to better capture the varying degrees (or dimensions) of drivel. Specifically, we aim to identify cross-narrative features of drivel as a step toward a broader investigation into its underlying properties beyond rigid narrative boundaries. In the present contribution, we propose six key dimensions for annotating drivel within a given text:

1. its distance from reality,
2. its linguistic and argumentative peculiarities,
3. claims to absoluteness (and overall handling of sources),
4. its suggestiveness,
5. its tendency to oversimplify complicated matters, and
6. the (apparent) emotionality of its author.

We present the development of a gold standard of 1,000 manually annotated German Telegram posts, including detailed guidelines and an analysis of inter-annotator agreement. Furthermore, we try to predict the overall drivel-like quality of a text from the annotated dimensions. In this context, we also analyse the correlations between these dimensions.[1]

2. Data

In 2020, as platforms like YouTube and Facebook intensified efforts to curb disinformation during the COVID-19 pandemic, skeptics, lockdown critics, and conspiracy theorists began migrating to alternative networks. Much of the text- and image-based discussion moved to Telegram – a minimally moderated messaging and microblogging platform – making its channels and groups key data sources for studying conspiracy theories (Lamberty et al., 2022; Holnburger et al., 2022).

2.1. Schwurpus: a corpus of conspiratorial talk

We use channels of prominent German COVID-19 conspiracy figures scraped via Telegram's export function (Heinrich et al., 2024). Since channels often interact through message forwarding, the corpus was expanded by iteratively including frequently mentioned channels with large follower counts, supplementing this with publicly available channel statistics. The final corpus – called "Schwurpus" – includes over 200 channels (followers ranging from a few thousands to over 300,000) and more than 100 public group chats from January 2020 to July 2022, totaling over 13 million posts and nearly 400 million tokens.

2.2. Sample

We drew a random sample from the Schwurpus for manual annotation. Only posts with 400 or more characters were

[1]The sample, guidelines, and adjudicated annotations can be found at https://github.com/fau-klue/infodemic.

considered. To ensure balanced representation and avoid bias towards highly active channels, the data were stratified by month and by channel frequency category. Two samples were initially drawn: the first consisted of approximately 1,000 posts and was originally used for the automatic detection of narratives related to the pandemic (Heinrich et al., 2024)[2], while the second comprised roughly 2,000 posts and included data from the entire Schwurpus. For the final dataset, both samples were merged and posts were sorted chronologically. To introduce narrative variation for initial testing, the first 100 posts were randomly sampled from the entire dataset.

In the present contribution, we present the results based on the first 1,000 posts of this combined dataset. The dataset contains posts dated between January 1, 2020, and July 29, 2022. The majority of posts were made during early to mid-2020, specifically between January and September. Only 83 posts were created after October 1, 2020. The posts originate from a total of 143 distinct channels. The channels with the highest number of posts are *evahermanoffiziell* (50 posts), *qglobalchange* (42 posts), *alternativemedien* (40 posts), *kulturstudio* (39 posts), and *oliverjanich* (35 posts). Texts are rather short, ranging from 56 to 1,024 tokens per post, with a median of 144 tokens and a mean of 208 tokens.

3. Annotation: dimensions of drivel

Since we assume that texts can be more or less drivel-like, it makes sense to visualise the opposite poles: at one end of the scale are incoherent texts full of far-fetched assertions without evidence, which are nonetheless presented in a tone of conviction; at the other end are fact-based, scientific texts with clean argumentation.

To characterise drivel, we defined six different dimensions. Student assistants rate texts on every dimension with values between 1 and 5. In addition, the overall drivel-like quality of a given text is also rated on the same scale (intuitively, without further instructions).

3.1. Guidelines

The guidelines to annotate posts include detailed descriptions of the six proposed dimensions of drivel, typical features, as well as fully annotated examples. Annotators were instructed to avoid middle values for inconspicuous texts, and to rate dimensions independently of each other.

Dimension 1: distance from reality. This category is concerned with how far a text departs from widely accepted reality, especially scientific consensus, by assessing the plausibility and number of assumptions required to believe its claims. The lowest rating indicates fact-based or experience-based content requiring no assumptions, while the highest indicates completely fabricated, fantastic or conspiratorial content requiring numerous implausible assumptions. Intermediate levels reflect increasing reliance on questionable premises, half-truths, unverifiable personal beliefs, or spiritual/religious claims with real-world implications.

[2]As this sample was drawn earlier, it does not include data from 2022.

Dimension 2: linguistic and argumentative peculiarities. This category assesses the linguistic and argumentative clarity of a text, focussing on whether conclusions logically follow from the stated premises. The lowest rating is reserved for clear, logical and coherent argumentation as well as for non-argumentative texts (e.g. purely social interactions). The highest rating indicates completely incoherent ramblings and texts where the argumentation is incomprehensible or outright missing (despite making claims). Key features include logical gaps, semantic incoherence, accumulations of grammatical or spelling mistakes, associative rather than logical reasoning, informal fallacies, personal attacks, and clickbait style. As these features become more frequent and disruptive, the rating increases.

Dimension 3: claim to absoluteness and handling of sources. This category evaluates how strongly authors assert their views – especially bold or controversial ones – and, to a lesser extent, how they handle sources. The lowest score reflects cautious, balanced language, hedging expressions, and a thoughtful, open engagement with evidence and alternative or opposing views. As scores increase, authors appear more convinced of their own perspective, use fewer qualifiers, and rely on anecdotal or selectively interpreted evidence. Higher scores indicate ideological rigidity, a lack of self-reflection, disregard or disparagement of differing views, manipulative use of sources and/or reliance on dubious ones. The highest score denotes an absolutist stance with unquestioned beliefs and missionary zeal.

Dimension 4: suggestiveness. This category assesses how much a text subtly tempts readers to draw unjustified conclusions without explicitly stating them. Key features include subtext, dogwhistling, implications, rhetorical questions, framing, loaded language, manipulative contrasts (splitting/black-and-white thinking), intentional use of fallacies, and emotionalisation. The lowest score is intended for texts free of suggestive elements. As the score rises, so does the presence of suggestive features, with the highest score indicating a clearly recognisable suggestive intention. Note that some of the features used here are also used in the annotation of framing and persuasion techniques (Piskorski et al., 2023).

Dimension 5: oversimplification. This category evaluates how much a text oversimplifies complex issues, particularly by presenting single causes for multifaceted problems or simply omitting key aspects. The lowest score indicates a nuanced, well-rounded presentation that respects a topic's inherent complexity and includes (nearly) all important factors. Increasing scores reflect texts beginning to generalise, ignore counterarguments and -evidence, and reduce explanations to fewer causes. At the highest level, the portrayal is extremely reductive – presenting only one factor as the sole explanation of one or even multiple issues, often in a way that distorts understanding and disregards complexity entirely.

Dimension 6: emotionality. This category assesses how emotional the author appears to be in their writing, be it angry, enthusiastic, despairing, or anxious. Key features of emotionality include words that explicitly refer to the author's emotional state, use of expressive emojis, dramatic

punctuation, and fully capitalised words or sentences. The lowest score indicates a neutral, objective tone with little to no emotional expression, whereas the highest is assigned to texts that are dominated by a very emotional or agitated tone.

3.2. Manual annotation

We employed a total of six annotators, with three participating in the first (finished) phase of the annotation process and another three in the second phase. In total, annotators 1–3 contributed 995 annotated posts (i.e. they annotated the complete sample). Annotators 4 and 5 provided an additional 300 annotated posts, and annotator 6 contributed a further 180 annotated posts – these annotations were predominantly executed for training the second batch of annotators, who are now annotating the next batch of the overall sample.

Figure 1 illustrates the score distributions for each annotator (note the varying y-axis scales, as the overall number of annotations differs across annotators). We observe that annotators 2 and 5 display a clear tendency to assign low scores, with annotator 5 in particular frequently opting for score 1. Annotator 6, by contrast, tends to assign scores 3 and 4 most of the time. In comparison, annotators 1, 3, and 4 exhibit a more uniform distribution across scores, though annotators 1 and 4 notably avoid the highest score (score 5). It is important to note that it is generally easier to reach high agreement when annotators consistently choose the middle of the scale (see below for inter-annotator agreement) but annotators were instructed to make clear-cut decisions and to avoid systematically choosing the middle scores.

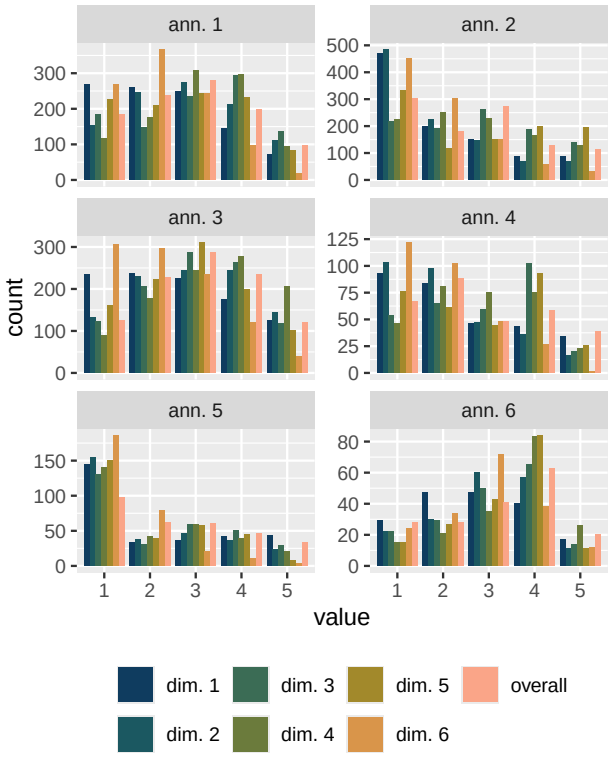

Figure 1: Distribution of annotation scores per annotator.

Additionally, manual adjudication was performed by two of the authors in collaboration with the annotators, focussing on the most challenging cases, i.e. those exhibiting the highest disagreement. This process was very slow, concentrating on difficult cases to ensure quality and improve understanding of each dimension. In total, 198 annotations were manually discussed and curated across all dimensions, plus 13 annotations regarding the overall drivel-like quality.

3.3. Inter-annotator agreement

Measuring agreement Inter-annotator agreement (IAA) scores were computed using Krippendorff's α for ordinal data (Landis and Koch, 1977), where the scores are treated as categories with inherent ranks. This measure can be calculated for pairs of annotators or for groups. Importantly, it accounts for the order of categories by employing a difference function d_{ij} that reflects the distance between categories i and j. In our analysis, we use the squared difference of ranks as the difference function:

$$d_{ij} = (r_i - r_j)^2$$

where r_i and r_j denote the ranks of categories i and j, respectively. The agreement coefficient is computed as

$$\alpha = 1 - \frac{D_o}{D_e}$$

where

$$D_o = \frac{\sum_{i,j} n_{ij} d_{ij}}{N}, \quad D_e = \frac{\sum_{i,j} n_i n_j d_{ij}}{N(N-1)}$$

represents the observed agreement and expected disagreement by chance, respectively. Here, n_i denotes the total number of times category i was assigned, N is the total number of assignments, and n_{ij} denotes the number of pairs of assignments where one annotation was given category i and the other category j.

Krippendorff's α ranges from -1.0, indicating perfectly discordant annotations, to $+1.0$, indicating perfect agreement. Negative values indicate agreement worse than chance. The interpretation of α values follows commonly accepted guidelines (Landis and Koch, 1977, 165): values between -1.0 and 0.0 indicate poor agreement; values greater than 0.0 up to 0.2 are considered slight; from 0.2 to 0.4 fair; from 0.4 to 0.6 moderate; from 0.6 to 0.8 substantial; and values above 0.8 up to 1.0 are interpreted as near-perfect agreement.

Overall agreement Agreement with the manually adjudicated data is generally very low, with most scores falling below zero. This is not unexpected, as we only adjudicated the most difficult cases, which naturally show higher disagreement. The highest individual agreement scores with the adjudicated data set were observed for annotation of the overall drivel-like quality and for dimension 1 (distance from reality), both reaching values above 0.6, which indicates substantial agreement.

Figure 2 visualises the distribution of IAA scores across all dimensions and for all subsets of annotators. Dimensions 1 (distance from reality) and overall drivel-like quality are clearly the most straightforward categories as measured by

the median of IAA scores – showing moderate to substantial median agreement levels and low variability. We observe moderate agreement for dimension 3 (claim to absoluteness and handling of sources). Moderate agreement, albeit with substantial variability, is also observed for dimension 6 (emotionality), dimension 2 (linguistic and argumentative peculiarities), and dimension 5 (oversimplification). In contrast, dimension 4 (suggestiveness) exhibits only fair agreement.

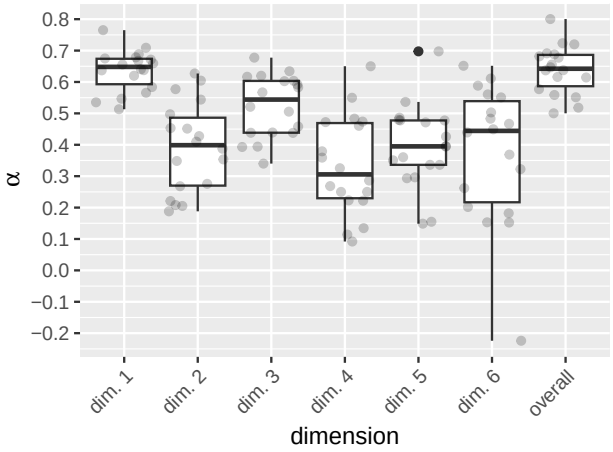

Figure 2: Distribution of agreement scores (Krippendorff's α) for each dimension.

Pairwise agreement Figure 3 presents average pairwise agreement scores for all dimensions. Overall, most annotator pairs demonstrate moderate agreement with one another, with a few notable exceptions. Annotators 2 and 4 exhibit substantial agreement on average, while annotators 3 and 5 as well as annotators 4 and 6 only achieve fair agreement. Agreement scores for the worst dimension (suggestiveness) range from .1 to .65, and from .51 to .77 for the best dimension (distance from reality). Interestingly, the agreement between annotators 2 and 4 remains exceptionally high – even substantial agreement for the most difficult dimension of suggestiveness – whereas annotators 3 and 5 drop down to slight agreement in this case.

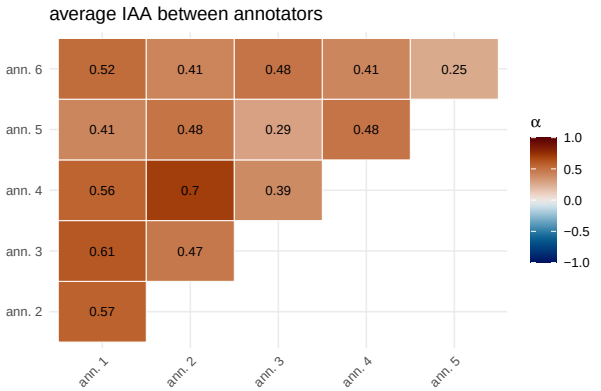

Figure 3: Average pairwise agreement scores across dimensions.

4. Prediction

How well can the dimensions predict the overall drivel scores of the annotated posts? To get a first impression using our annotated sample, we took the adjudicated values, added mean values of all annotators for non-adjudicated dimensions, and used the resulting dataset for simple multiple linear regression analyses.

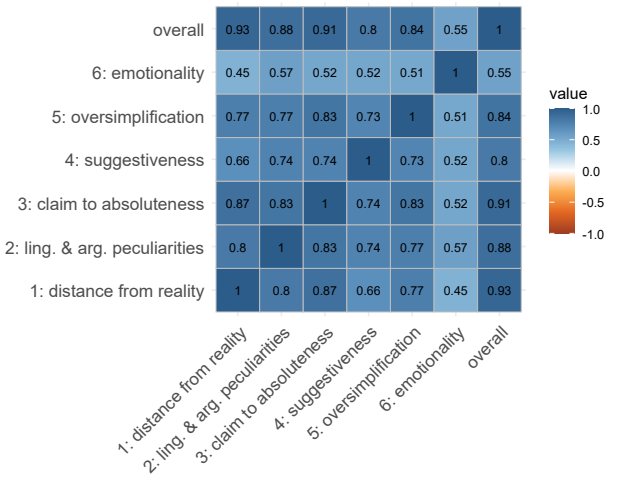

Figure 4: Pearson correlation coefficients between dimensions and overall scores.

Figure 4 shows the correlation coefficients between the individual variables, while Table 1 shows results from two different regression models. As can be clearly seen from the full model, *distance from reality* appears to be the single most important predictor (whereas *emotionality* does not contribute much). Even the second model, which includes only predictors that can be approximated by linguistic features alone (rather than, e.g., world knowledge), retains much explanatory power.

Table 1: Full regression model (left) and model with reduced number of predictors (right).

	Dependent variable:	
	'overall'	
	model 1	model 2
'1: distance from reality'	0.469*** (0.016)	
'2: ling. & arg. peculiarities'	0.163*** (0.017)	0.406*** (0.023)
'3: claim to absoluteness'	0.156*** (0.019)	0.608*** (0.020)
'4: suggestiveness'	0.201*** (0.014)	
'5: oversimplification'	0.087*** (0.015)	
'6: emotionality'	0.042*** (0.012)	0.043** (0.018)
Constant	−0.174*** (0.028)	−0.151*** (0.039)
Observations	995	995
R^2	0.945	0.879
Adjusted R^2	0.945	0.879
Residual Std. Error	0.267 (df = 988)	0.396 (df = 991)
F Statistic	2,843.781*** (df = 6; 988)	2,398.010*** (df = 3; 991)
Note:		*p<0.1; **p<0.05; ***p<0.01

5. Conclusion

We presented an analysis of drivel found in German Telegram in terms of six distinct dimensions, based on a sample of 1,000 posts. Consistent manual annotation is a difficult endeavour, yet we can show moderate to substantial agreement between annotators. We will continue our effort to provide high-quality annotation for a larger sample. We hope to eventually provide a rich resource both for computational linguistics (e.g. automatic prediction of individual dimensions and overall drivel-like quality from text) and for corpus linguistics (potentially providing insight into how different dimensions manifest linguistically).

We also showed that the prediction of overall drivel-like quality from the six dimensions is straightforward. A simple linear regression model shows that posts are especially likely to be considered drivel when they appear distant from reality and when authors strongly assert their views. Future work will comprise further analysis of the associations between different dimensions.

A key limitation of the present study is the ambiguity and overlap among the six annotation dimensions. In particular, dimension 2 (linguistic and argumentative peculiarities) conflates several distinct elements – including spelling errors, coherence violations, and argumentative inconsistencies – into a single category. This lack of clear separation between formal, semantic, and logical features makes the development of precise annotation guidelines difficult, which in turn contributes to relatively low inter-annotator agreement and ultimately makes the dimension difficult to interpret. Future iterations of the study will aim to refine the dimensional structure to address these issues.

Acknowledgements This research has been partially funded by the German Research Foundation (DFG), project no. 466328567.

6. References

Douglas, K. M., Uscinski, J. E., Sutton, R. M., Cichocka, A., Nefes, T., Ang, C. S., and Deravi, F. (2019). Understanding conspiracy theories. *Political Psychology*, 40(S1):7.

Heinrich, P., Blombach, A., Doan Dang, B. M., Zilio, L., Havenstein, L., Dykes, N., Evert, S., and Schäfer, F. (2024). Automatic identification of COVID-19-related conspiracy narratives in German telegram channels and chats. In Nicoletta Calzolari, et al., editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1932–1943, Torino, Italia, May. ELRA and ICCL.

Holnburger, J., Goedeke Tort, M., and Lamberty, P. (2022). Q vadis? Zur Verbreitung von QAnon im deutschsprachigen Raum.

Lamberty, P., Holnburger, J., and Goedeke Tort, M. (2022). Das Protestpotential während der COVID-19-Pandemie.

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

Piskorski, J., Stefanovitch, N., Nikolaidis, N., Da San Martino, G., and Nakov, P. (2023). Multilingual multifaceted understanding of online news in terms of genre, framing, and persuasion techniques. In Anna Rogers, et al., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3001–3022, Toronto, Canada. Association for Computational Linguistics.

Piskorski, J., Mahmoud, T., Nikolaidis, N., Campos, R., Jorge, A., Dimitrov, D., Silvano, P., Yangarber, R., Sharma, S., Chakraborty, T., Guimarães, N., Sartori, E., Stefanovitch, N., Xie, Z., Nakov, P., and Da San Martino, G. (2025). SemEval-2025 task 10: Multilingual characterization and extraction of narratives from online news. In *Proceedings of the 19th International Workshop on Semantic Evaluation*, SemEval 2025, Vienna, Austria, July.