Extending a Parliamentary Corpus with MPs' Tweets: Automatic Annotation and Evaluation Using TTLABTWEETCORPUS

Mevlüt Bagci and Ali Abusaleh and Daniel Baumartz and Alexander Mehler and Giuseppe Abrami and Maxim Konca

Text Technology Lab, Institute of Computer Science and Mathematics,
Goethe University Frankfurt, Robert-Mayer-Strasse 10,
60325 Frankfurt am Main, Germany
{bagci, a.abusaleh, baumartz, mehler, abrami, konca}@em.uni-frankfurt.de

Abstract

Social media is central to modern politics, reflecting politicians' ideologies and enabling outreach to younger generations. We present TTLABTWEETCORPUS, a multilingual X tweet corpus linking politicians' social media discourse to the German parliamentary corpus GERPARCOR, enabling comparisons between online communication and parliamentary debates. TTLABTWEETCORPUS contains 39 546 tweets, including 19 056 tweets with media content, collected using our newly developed general-purpose X data collection tool, TTLABTWEETCRAWLER. We used nine text models and one vision-language model (VLM) to annotate TTLABTWEETCORPUS for emotion, sentiment, and topic, and evaluated them against a manually annotated subset. We provide TTLABTWEETCORPUS with automatic text- and media-based annotations validated against human labels, together with the accompanying TTLABTWEETCRAWLER tool. Our analysis shows that the models are mutually predictable. We show that annotators prefer VLM-based labels, suggesting multimodal representations better align with human interpretation.

1 Introduction

Social media platforms such as X, TikTok, and Instagram have become increasingly influential in society. Political content on these platforms can shape public opinion and affect electoral outcomes (Olaniran and Williams, 2020). As such content is often multimodal, political communication analysis requires corpora that integrate social media with political datasets. We address this gap by introducing TTLABTWEETCORPUS, which contains over 39 546 tweets, including 19 056 with media information. Over 15 000 of these tweets are from German politicians during the election period from 2024 to 2025, and all are linked to GERPARCOR (Abrami et al., 2022, 2024).

Constructing TTLABTWEETCORPUS, a multimodal corpus primarily comprising political tweets, involved three main steps: collecting tweets and media, linking tweets to German parliamentary speeches, and generating automatic annotations and evaluating them through manual annotation. For data collection, we developed TTLABTWEETCRAWLER, a general-purpose X data collection tool with two methods: a user-based method that does not guarantee media content and a search-based method that targets tweets with media information.

Additionally, we used nine text-based models spanning three types of classification, within the Docker Unified UIMA Interface (DUUI; Leonhardt et al. (2023)) to annotate all 33 921 nonempty tweets. The models were deployed as containerized DUUI components, allowing their scalable orchestration in a Kubernetes environment (Abrami et al., 2025). DUUI additionally supports the distributed processing of multimodal data (Bundan et al., 2025).

For each model, we trained a classifier using annotations from the remaining models as input features and optimized it via evolutionary search[1]. We analyze dependencies among the model outputs to identify redundant annotations and determine whether comparable performance can be achieved with fewer features. In addition to evolutionary search, we used *SHAP* (SHapley Additive exPlanations) (Lundberg and Lee, 2017) to analyze the impact of each input feature on the prediction model. Our evaluation shows that target model outputs can be predicted from annotations generated by the remaining models. Furthermore, we annotated all three classification types using an LLM capable of processing media information, with 18 703 processable tweets.

We manually annotated a sample of TTLAB-

[1]We use the implementation by Fortin et al. (2012)

TWEETCORPUS in order to assess and compare the performance of text- and media-based annotation tools. Our findings show that tweet classification should not rely on text alone, but also incorporate relevant media information. Due to data rights restrictions, we release tweet IDs in TTLABTWEETCORPUS, but not the tweets or their media information. In addition, the corpus includes user IDs, links to GERPARCOR, and all automatic and manual annotations. Our code and corpus are available on GitHub under the AGPL license[2].

2 Related Work

For more than 10 years, researchers have widely used corpora based on tweets from the X platform (McCreadie et al., 2012; Camacho-Collados et al., 2022; Blakey, 2024). Examples include COVID-19 research, such as the X corpus by Naseem et al. (2021), which contains 90 000 tweets from February to March 2020. The tweets are automatically annotated with sentiment polarities, and both the texts and annotations are available for download. More corpora are available from Müller and Salathé (2019), Lopez and Gallemore (2021), Hayawi et al. (2022), and Langguth et al. (2023).

Gonzalez (2024) presents an annotated corpus of 112 690 English tweets from 26 news agencies and 27 individuals, from 2009 to 2022. The automatically generated annotations are visualized in an interactive web application. However, the authors released neither information on the selected accounts nor the tweet texts or corresponding IDs, which hinders reproducibility. Derczynski et al. (2016) developed an X corpus for named entity research with approximately 156 000 tokens from tweets posted between 2009 and 2014. Annotated by experts and crowdsourcing, the corpus is available for download and includes texts and annotations. Additional corpora include 52 million manually labeled crisis-related tweets across 19 crises by Imran et al. (2016) and 97 million tweets by Petrović et al. (2010), although the latter is no longer available online.

Barbieri et al. (2020) proposed a unified X framework to simplify the evaluation and benchmarking of tweet classification tasks. This framework collects datasets for seven tasks, including hate speech detection and sentiment analysis. The tweets are labeled according to the task and based on existing X datasets created by Mohammad et al. (2018), Barbieri et al. (2018), Van Hee et al. (2018), Basile et al. (2019), Zampieri et al. (2019), Rosenthal et al. (2017), and Mohammad et al. (2016).

Although prior corpora are mostly text-based, work on media content in tweets (Thakkar et al., 2024; Konyspay et al., 2025; Paul et al., 2025) highlights the need for multimedia corpora. This paper introduces the X Corpus TTLABTWEETCORPUS, which contains both media and text data. Similarly, this corpus by Chen and Zou (2023) is a dataset of 800 000 tweets with AI-generated images that were collected between January 2021 and March 2023. Although the release includes generated captions and metadata, including tweet IDs, the tweet texts are not available. Balasubramanian et al. (2025) built an X corpus from tweets related to the 2024 U.S. presidential election, including public tweets and additional interaction and media information. Grimminger and Klinger (2021) published 3 000 tweets from the 2020 U.S. elections, manually annotated for stance and hate speech detection, releasing tweet texts but not IDs. In this paper, we present an X corpus of tweets from German politicians to enable analyses of political communication. Likewise, De Smedt and Jaki (2018) released a corpus of 125 000 tweets and 4 000 linked images. This corpus is no longer publicly available. Weissenbacher and Kruschwitz (2024) have released an annotated dataset of 1 250 tweet IDs about German MPs for research on offensive language and hate speech. Castanho Silva and Proksch (2022) formed a new corpus by merging the tweets of EU national parliament members (Castanho Silva and Proksch, 2021) with their parliamentary speech records Rauh and Schwalbach (2020) from 2018. Further unpublished X corpora from political research include Conover et al. (2011); Serrano et al. (2019); Breeze (2020); Schmidt et al. (2022); Ghoraba (2023); Hellwig et al. (2023).

In contrast to previous resources, TTLABTWEETCORPUS covers tweets from 2024 to 2025 and includes their associated media content.

3 Data

TTLABTWEETCORPUS focuses on two main aspects of tweets from X. First, it includes user tweets, with particular attention to German politicians during the 2025 election period. By link-

[2] https://github.com/texttechnologylab/TTLabTweetCorpus

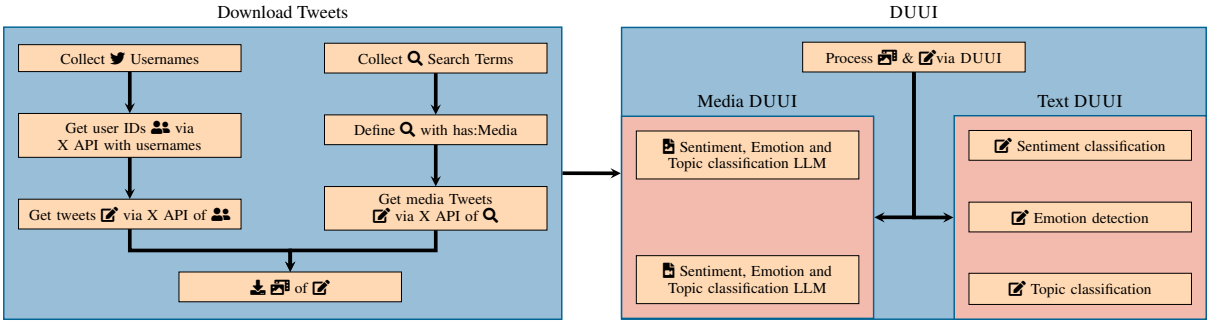

Figure 1: Building TTLABTWEETCORPUS: Overview of the process of the crawling of tweets and media data and the annotation using DUUI.

ing official speeches with social media activity, it supports the study of political communication, especially during election campaigns (Olaniran and Williams, 2020). Second, TTLABTWEETCORPUS includes media content embedded in tweets, adding multimodal context to the textual data. Within TTLABTWEETCRAWLER, we implemented the search-based method to systematically acquire tweets with media content. This method complements the tweets collected using the user-based method with multimodal context for political communication.

All tweets in TTLABTWEETCORPUS were collected from X using the *X API v2*[3]. To facilitate reproducibility, the complete dataset can be reconstructed through our TTLABTWEETCRAWLER. Figure 1 illustrates the full workflow, from data acquisition to the classification of tweets. As previously mentioned, our TTLABTWEETCRAWLER retrieves not only the textual content and standard metadata provided by the X API, but also the associated images and videos. The download process consists of two steps: first, tweets are retrieved; then, each tweet's media information is extracted to download the corresponding images or videos. We use two collection methods: the user-based method and the search-based method, with the latter targeting tweets containing media content. Figure 3 shows the number of collected tweets and how many contain images or videos.

For the user-based method, we compiled a list of X usernames belonging to German politicians. To do so, we extracted politicians' names and their listed X accounts from the publicly accessible websites of the political parties included in this study and manually verified and corrected the information where necessary (SPD-Bundestagsfraktion, 2026; CDU/CSU-Fraktion, 2026; AfD-Fraktion im Deutschen Bundestag, 2026; Fraktion der Freien Demokraten im Deutschen Bundestag, 2026; Fraktion Die Linke im Bundestag, 2026; Bundestagsfraktion Bündnis 90/Die Grünen, 2026; BSW-Gruppe im Bundestag, 2026). The official party websites were identified via the German Bundestag website (Deutscher Bundestag, 2026). The selected parties were represented in the German Bundestag both prior to the coalition's collapse in 2024 and following the 2025 federal election. In addition, BSW was included as it participated in the election (Belousova et al., 2025).

We used the X API to retrieve user IDs and download all timeline tweets from November 6, 2024 to February 23, 2025, spanning the day after the German coalition government collapsed through the end of the 2025 election. In total, we collected 15 175 tweets from 69 politicians across seven parties. Figure 2 shows tweet counts by party on the left and the top 10 individual politicians by tweet count on the right. In TTLABTWEETCORPUS, "Die Grünen" posted the most tweets, while "Dagmar Schmidt" was the most active individual politician.

As noted above, TTLABTWEETCORPUS extends politicians' speeches with their social media tweets and associated media content. We added further information by selecting stenographically recorded speeches of the individual politicians from GERPARCOR, limited to the 20th and 21st legislative periods in accordance with the dataset. The selected speeches do not include any comments from the plenary session; however, they do include the corresponding speaker and their parliamentary group affiliation. We used two methods to link tweets to GERPARCOR. First, X usernames were

[3] https://docs.x.com/x-api/introduction. Last accessed: 2026-08-07

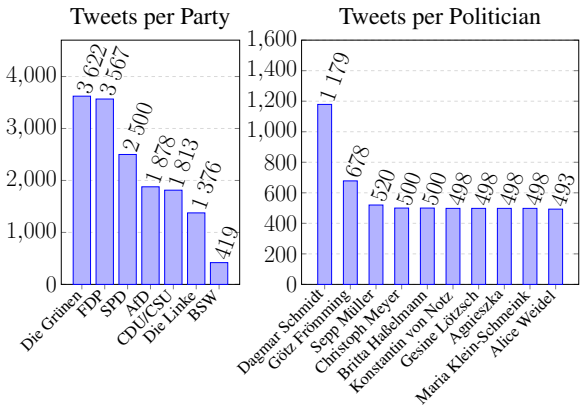

Figure 2: Number of tweets: party (left); top 10 individual politicians (right) in X dataset

matched to GERPARCOR politicians by name identity, yielding one-to-one mappings where accounts could be clearly assigned to known members of parliament. Second, we established semantic links between tweets and parliamentary speeches using cosine similarity. The speeches in GERPARCOR were segmented into sentences using *spaCy* (Honnibal et al., 2020). Sentence embeddings were generated for the tweets and speech sentences using `paraphrase-multilingual-MiniLM-L12-v2` (Reimers and Gurevych, 2019), selected for its multilingual support and suitability for semantic sentence similarity. For each tweet–speech pair, we calculated cosine similarity between the tweet embedding and every sentence embedding of the speech and averaged the resulting sentence-level scores. Users can define an application-specific similarity threshold or disable threshold-based filtering entirely. When a threshold is applied, pairs exceeding it are considered semantically linked. This procedure permits many-to-many relations, as a tweet may be linked to multiple speeches and a speech to multiple tweets.

After downloading, we used the URLs in the tweets' media information to retrieve associated images or videos. However, the user-based method does not guarantee media content. In practice, it yielded 22 072 tweets, of which 4 736 included media content. Therefore, we additionally employed the search-based method to target tweets containing media information. The search-based method requires search terms, which we generated by extracting the top 100 hashtags via the X API for a predefined list of German- and English-speaking countries[4]. We assign a weight of 100 to the top-ranked hashtag, 1 to the lowest-ranked hashtag and decrement the weight by one for each subsequent rank. Since the search-based method targets tweets containing media rather than tweets by specific German politicians, the retrieved tweets cannot be linked to GERPARCOR by name identity and are therefore linked using the semantic similarity procedure described above. The top 100 hashtags from each country are then combined and sorted by weight. For each search term, tweets are retrieved using the X API filter `has:media_link -is:retweet` to collect only tweets containing media links and exclude retweets. As with the user-based method, the corresponding media files are retrieved after the tweets have been downloaded via their media links.

Figure 3 summarizes the contributions of the two collection methods to TTLABTWEETCORPUS. The user-based method yielded 22 072 tweets, including 2 464 tweets with images and 2 272 with videos. The search-based method contributed 17 474 tweets, including 12 180 tweets with images and 2 140 with videos. Overall, TTLABTWEETCORPUS contains 39 546 tweets, of which 19 056 contained media content: 14 644 contained images and 4 412 contained videos. Of the tweets in TTLABTWEETCORPUS, 14 756 tweets were classified by both text- and media-based models. Another 3 947 tweets were classified only by the media-based models, as they contained no textual content after cleaning and filtering. Conversely, 19 165 tweets were classified only by the text-based models, as they either lacked media or contained broken media links. Finally, 1 678 tweets were excluded because they had neither text after cleaning nor valid media. The discrepancy between collected tweets and tweets with media arises because the implementation also downloads tweets connected to the originally retrieved tweets. Thus, the search-based method ensures media content only for the originally retrieved tweets, not necessarily for connected tweets.

4 Automatic Annotation Pipeline

TTLABTWEETCORPUS contains text and media that require separate processing: some models are text-only, while a subset supports multimodal input for images and videos in our pipeline. The right side of Figure 1 shows the pipeline used to process both text- and media-based tweets. All tweets, regardless of modality, were processed with DUUI,

[4]Countries and hashtags used are listed at GitHub

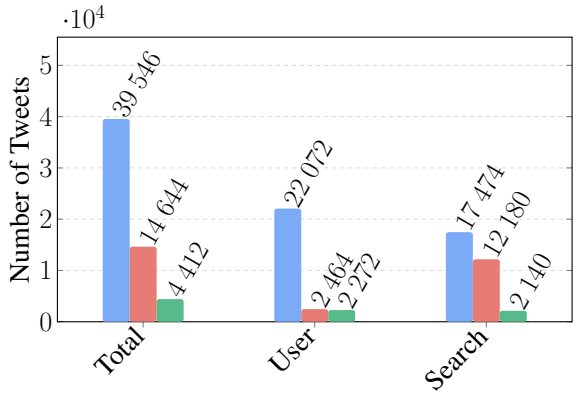

Figure 3: Numbers of tweets collected for TTLABTWEETCORPUS via TTLABTWEETCRAWLER using the user-based and search-based methods (Tweets: ■; Images: ■; Videos: ■).

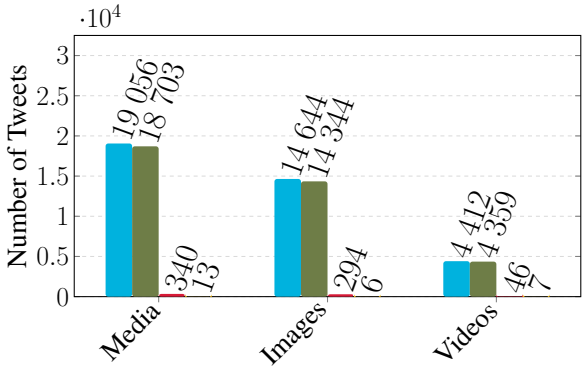

Figure 4: Bar chart summarizing the status of media file processing. Categories include processed, broken, and remaining files for images, videos, and totals (Tweets: ■; Processed: ■; Base broken: ■; Runtime Broken: ■)

Name	Citation
E1	Shivshankar (02shanky) (2023)
E2	Bertolini et al. (2024)
E3	Widmann and Wich (2023)
T1	Uslu et al. (2018)
T2	Kuzman and Ljubešić (2025)
T3	Antypas et al. (2024)
S1	CitizenLab (CitizenLabDotCo) (2022)
S2	Schmid (2022)
S3	Barbieri et al. (2022)
L1	Bai et al. (2025)

Table 1: Overview of the sentence-level classification tasks and the corresponding models used (E=Emotion, T=Topic, S=Sentiment; L indicates LLM-based classification for the respective task).

which integrates NLP and LLM-based models for different tasks via Docker containers.

4.1 Text-Based Analysis

All textual tweets were cleaned by removing URLs, hashtags, user mentions, and emojis, and all empty tweets were excluded. After cleaning 39 546 entries, 33 921 tweets contained valid textual information. The remaining tweets were segmented into sentences using *spaCy* for classification. Three sentence-level classification tasks—emotion, topic, and sentiment detection—were applied, each using three models. The selected models for each task are listed in Table 1. The models differ in architecture and training domain. E1, E2, T2, and S1 are based on `XLM-RoBERTa` variants; E3 uses `mDeBERTa`, T1 `fastText`, S2 `multilingual DistilBERT`, and S3 `XLM-T`. T3 is a multilingual social-media classifier, while L1 uses the multimodal `Qwen2.5-VL-3B` model. Their training domains include Twitter, political communication, news, and general multilingual text.[5] For each tweet–model combination, category scores were averaged across all sentences, normalized using softmax, and the category with the highest resulting probability was selected as the tweet-level label. The models retained their respective task-specific label sets; further details are provided in Section 6.

4.2 Media-Based Analysis

The media data were also processed via DUUI-Multimodal (DUUI-MM; Abusaleh (2025)) using the model listed under L1 in Table 1. As previously mentioned, TTLABTWEETCORPUS contains a total of 19 056 tweets with media content. A total of 18 703 media files were successfully processed, consisting of 14 344 image-based tweets and 4 359 video-based tweets. Additionally, 340 media tweet files were found to be broken during retrieval: 294 images and 46 videos. Furthermore, 13 files failed during runtime due to corrupted data: 6 images and 7 videos. Image-processing failures were mainly caused by model errors with certain image encodings, particularly malformed base64 strings or data unsupported by standard decoders. Video failures mostly resulted from corrupted or incomplete stream files that could not be parsed. Figure 4 summarizes these statistics and enables a visual comparison across media types.

[5] Detailed model descriptions, label inventories, and implementation links are available at GitHub.

4.2.1 Prompt Construction for Multimodal Media Classification

The prompt is designed to replicate text-based classification behavior within a multimodal reasoning setting. It is instantiated through the L1 model, which serves as the vision-language backbone for multimodal inference. The prompt supports three interdependent classification dimensions: emotion recognition, sentiment polarity, and topic categorization based on the Dewey Decimal Classification (DDC) (OCLC, 2025). Emotion recognition uses five affective categories (*anger*, *sadness*, *apprehension*, *confusion*, *happiness*) from Model E2; sentiment analysis follows the Model S1 polarity labels (*positive*, *neutral*, *negative*); and topic classification uses the ten main DDC classes to ensure compatibility with knowledge organization standards. The prompt enforces three principles: *Modularity*, by separating subtasks with distinct label spaces and output schemas; *Interpretability*, by requiring evidence-based explanations grounded in visual cues; and *Structure*, by producing valid JSON with confidence distributions summing to one. These constraints ensure the prompt is both expressive and evaluable. A high-level abstraction of the full prompt is shown in Listing 1[6] in the Appendix A.

4.2.2 Media-Based Classification Workflow

After prompt execution, the VLM processes the structured input and generates a single output containing predictions for each subtask. A post-processing layer then applies consistency checks, extracts JSON, performs semantic validation, and aligns the predictions with the controlled label sets. Finally, the system produces a unified annotation containing emotion, sentiment, and topic information.

5 Manual Annotation

We evaluated the outputs of the text- and media-based models using manually assigned labels and annotators' preferences between the two model types. Seven professionals annotated 103 tweets, resulting in 721 tweet-level annotation instances. The annotators were research assistants, PhD candidates, and postdoctoral researchers from linguistics and computer science. The annotation was conducted using a web application based on TEXTANNOTATOR (Abrami et al., 2020). For each tweet, annotators were shown its text, any accompanying image or video, and the classification outputs of the text- and media-based models. Each annotator made six decisions per tweet: they assigned one topic, one sentiment polarity, and one emotion using the label sets of models T1, S1, and E2 defined in Section 4.2.1, and selected the better-fitting model output for each of the three tasks. For each model comparison, annotators could select the text-based model, the media-based model, both models, or neither model. Thus, manual labeling and model-preference evaluation were conducted within the same annotation session.

We report inter-rater reliability based on Krippendorff's alpha (Krippendorff, 1970) and Fleiss's kappa (Fleiss, 1971)[7] in Table 2. The results are mixed: annotators agree on topic classification (82%), but reliability for sentiment and emotion drops to 45–55%, indicating difficulties in assigning these features to tweets, as also reported by e.g., Bostan et al. (2020). For the manually assigned labels, agreement was highest for topic classification ($\alpha = 0.82$, $\kappa = 0.82$), followed by sentiment ($\alpha = 0.54$, $\kappa = 0.54$) and emotion ($\alpha = 0.47$, $\kappa = 0.46$). Agreement on model preference was lower, with $\alpha = 0.66$ and $\kappa = 0.66$ for topic, $\alpha = 0.33$ and $\kappa = 0.33$ for sentiment, and $\alpha = 0.34$ and $\kappa = 0.33$ for emotion. These results indicate that topic labels were assigned comparatively consistently, whereas sentiment and emotion labels, as well as model preferences, showed lower inter-rater reliability, consistent with the challenges reported by Bostan et al. (2020). For topic classification, annotators preferred the media-based model in 49% of the decisions and the text-based model in 23%, while both models were selected in 19% and neither model in 9%. For sentiment, the corresponding proportions were 42%, 34%, 14%, and 11%; for emotion, they were 41%, 33%, 16%, and 10%, respectively. For topic classification, they preferred the media model by a larger margin (49% vs. 23%).

To derive one reference label per tweet and classification task, we aggregated the seven human-assigned labels using majority vote (MV) and Dawid–Skene (DS) aggregation (Dawid and Skene, 1979)[8] MV and DS are alternative methods for constructing reference labels and are not classification models themselves. We evaluated the

[6] Code and full prompt are available at GitHub

[7] We use the methods of Castro (2017) and Seabold and Perktold (2010) to calculate inter-rater reliability.

[8] We use the implementation by Ustalov et al. (2024).

predictions of the media- and text-based models against the MV- and DS-derived reference labels using Macro F1; the model predictions were not evaluated against each annotator's labels individually. Table 3 reports the resulting scores for both reference-label aggregation methods. For topic classification, the media-based model achieved Macro F1 scores of 0.56 and 0.48 against the MV- and DS-derived reference labels, respectively, whereas the text-based model achieved 0.10 and 0.13. For sentiment, the text-based model outperformed the media-based model, obtaining 0.66 and 0.68 against the MV- and DS-derived reference labels, respectively, compared with 0.58 and 0.59 for the media-based model. The text-based model also performed better for emotion, obtaining a score of 0.58 and 0.52, compared with 0.45 against MV and 0.44 against DS for the media-based model. Overall, the media-based model performed better for topic classification, whereas the text-based model performed better for sentiment and emotion. As a chance baseline, we randomly assigned reference labels to the tweets and recomputed the Macro F1 scores over five runs. The resulting mean scores were substantially lower for topic classification and generally lower for sentiment and emotion, as shown by the RND rows in Table 3. This comparison indicates that the models align more closely with the human-derived reference labels than with randomly assigned labels. However, performance above the random baseline does not by itself demonstrate high annotation reliability, particularly given the lower inter-rater reliability for sentiment and emotion. At the same time, the overall model performance, especially for topic classification, reveals the intrinsic difficulty of interpreting multimedia content through textual features alone.

This finding reinforces the assumption that tweets, as highly multimodal entities, require joint modeling of text and media components to capture their full semantic and affective meaning.

6 Cross-Model Annotation Analysis

In Section 4.1, all cleaned textual tweets ($\mathcal{X}$) were processed using various classification models $\mathcal{T} = \{t_1, \ldots, t_9\}$. To examine redundancy and dependencies among the automatic annotations, we test the following hypothesis: **H1:** The output of a target model t_i can be predicted from the outputs of the remaining models $\mathcal{R} = \mathcal{T} \setminus \{t_i\}$. This post-hoc analysis examines outputs from models with different architectures, training domains, and label inventories and tests whether fewer models or features can retain comparable predictive performance, rather than comparing standalone model quality.

To test H1, we generate annotation-output vectors for each model t_i and tweet x ($x \in \mathcal{X}$). This post-hoc analysis identifies redundant model outputs and tests whether fewer models or features can retain comparable predictive performance. For each model, we compute the mean score of each target category across all sentences in x and normalize these scores with a softmax function so that they sum to one. Using this procedure, we create a vector for each text x, based on the remaining $\mathcal{R}$ models corresponding to model t_i. The predicted target label for each model is defined as the category with the highest average probability. We repeat this procedure for each model and tweet. Labels represented by fewer than 50 tweets are removed. The dataset is split into two parts: 80% for training and 20% for testing. We use a

	MP		GD	
Task	$\alpha \uparrow$	$\kappa \uparrow$	$\alpha \uparrow$	$\kappa \uparrow$
Topic	0.66	0.66	0.82	0.82
Sentiment	0.33	0.33	0.54	0.54
Emotion	0.34	0.33	0.47	0.46

Table 2: Inter-rater reliability (Krippendorff's α, Fleiss' κ) for model preference (MP) and gold-data generation (GD).

	Agg.	F1 Media (#S)	F1 Text (#S)
Topic	MV	0.56 (103)	0.10 (97)
Sentiment	MV	0.58 (103)	0.66 (97)
Emotion	MV	0.45 (103)	0.58 (97)
Topic	DS	0.48 (103)	0.13 (97)
Sentiment	DS	0.59 (103)	0.68 (97)
Emotion	DS	0.44 (103)	0.52 (97)
Topic	RND	0.11 (103)	0.05 (97)
Sentiment	RND	0.32 (103)	0.32 (97)
Emotion	RND	0.18 (103)	0.17 (97)

Table 3: Macro F1 of the media- and text-based models evaluated against reference labels derived from the human annotations using majority vote (MV) and Dawid–Skene aggregation (DS). RND denotes the mean Macro F1 over five runs using randomly assigned reference labels; #S denotes the number of evaluated samples.

Random Forest classifier to test H1. This study involves a total of nine models, and nine prediction models are trained using this approach. All models are optimized with evolutionary search to show that equal or nearly equivalent Macro F1 can be achieved with substantially fewer features. The evolutionary search was run for 200 generations, with each candidate feature subset evaluated using a Random Forest classifier implemented in `scikit-learn` v1.1.2 (Pedregosa et al., 2011) (`n_estimators=100`, `min_samples_leaf=5`, `class_weight="balanced"`; with a fixed random seed and all other hyperparameters set to their default values).

Figure 5 shows Macro F1 before and after feature selection, along with the number of target categories per model. Nearly half of the features were unnecessary for achieving optimal Macro F1, with models using about 55% of available features on average. The average Macro F1 across models is 65% (34%–99%), reaching 99% for all sentiment classification types. This likely reflects the small number of labels to predict. Spearman's rank correlation (Zar, 2014) shows a strong negative association between Macro F1 and the number of labels ($\rho = -0.85$, $p < 0.01$), indicating lower performance with more labels. Sentiment labels are more balanced across models than topic and emotion labels.

Figure 6 shows, for each trained classification model, the proportion of selected features originating from each feature model; higher bars indicate higher proportions. E1 consistently ranks last, suggesting its feature parameters are not useful after feature selection. E3 performs best, with an average rank of 3.25, followed by S2, with an average rank of 3.75; the remaining models occupy middle ranks on average. Within this cross-model prediction setup, features derived from E1 were selected least often, whereas E3 achieved the best average rank of 3.25, followed by S2 with 3.75. These ranks reflect the utility of a model's outputs for predicting the other models, not its standalone annotation quality. Differences may result from the models' training domains, label definitions, class distributions, and correlations with the other outputs rather than from the numbers of features or labels alone.

The colors indicate the mean *SHAP* influence per feature model, from low (blue) to high (yellow). For comparability, *SHAP* was computed using the same training and test splits as in the evolutionary search. We computed mean *SHAP* values across all labels for each feature, capturing each feature parameter's contribution to the target labels. This aggregated score is visualized in Figure 6, and the sum of *SHAP* values reflects each feature's total influence on the prediction model. It also suggests a correlation between the features selected through evolutionary search and their *SHAP*-derived influence on the model.

Spearman rank correlation confirms this observation using selection proportion and aggregated *SHAP* influence as inputs ($\rho = 0.36$, $p < 0.01$). This indicates a statistically significant but slight positive monotonic relationship between feature selection frequency and *SHAP* influence. The overlap between the top 20 individual feature parameters per model, ranked by evolutionary selection frequency and by *SHAP* influence, is exactly 40%. This further suggests that evolutionary search successfully identifies features that meaningfully affect model predictions.

The analysis reveals dependencies and redundancy among the automatic annotations in TTLABTWEETCORPUS and shows that comparable predictive performance can be retained with fewer features. It thereby supports more efficient annotation pipelines and joint analyses of topic, sentiment, and emotional framing in political communication.

7 Conclusion

TTLABTWEETCORPUS offers more than just text data: it comprises 15 175 tweets from German political figures, which are linked to GERPARCOR. TTLABTWEETCRAWLER supports corpus extension through two collection methods: the user-based method yields more text-only tweets, whereas the search-based method mainly retrieves tweets containing media. Together, both processes produce 39 546 tweets, of which 19 056 include media information and 18 703 are processable. We used nine text-based models, which cover three types of classification, to analyze all 33 921 tweets containing text after cleaning. Topic, emotion, and sentiment can be predicted in relation to each other, and evolutionary feature selection shows that comparable or higher Macro F1 can be achieved with fewer features. We also introduced a VLM-based method for automatically annotating media content, such as images and videos, using a single prompt for the same three classification types. A small-scale manual evaluation shows

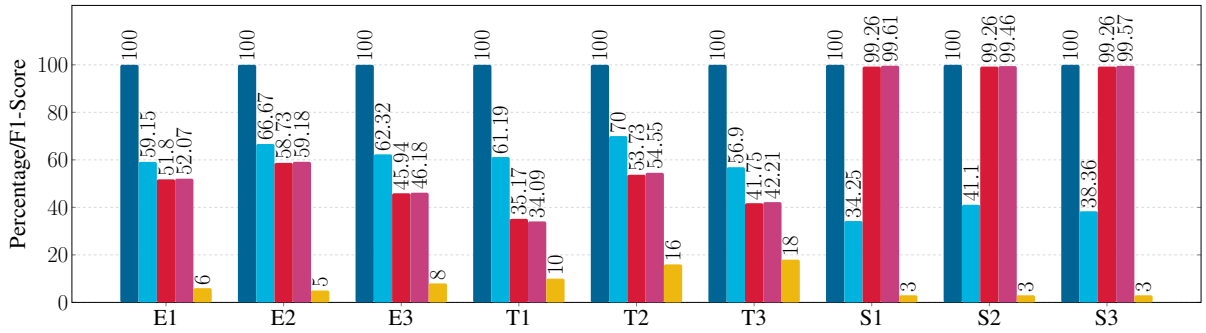

Figure 5: Performance results (Macro F1 and number of features) of the 3 classification tasks using an evolutionary search for optimization of feature selection (Features: ■; Selected Features: ■; Macro F1: ■; Optimal Macro F1: ■; Labels: ■)

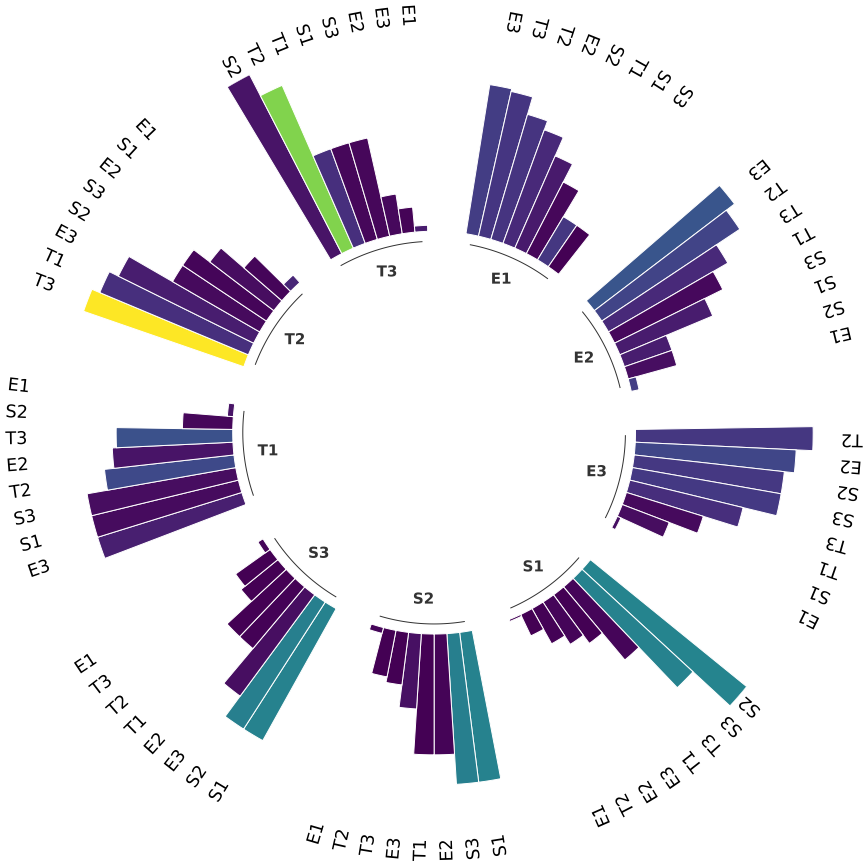

Figure 6: *SHAP* analysis of feature importance vs. the percentage of selected features for the best evolutionary-search model, with *SHAP* impact shown by color (low: blue; high: yellow)

that annotators more often preferred the media-based outputs. However, the media-based model performed better for topic classification, whereas the text-based model achieved higher Macro F1 scores for sentiment and emotion. However, further annotation is required because the current manually labeled dataset is limited. Nevertheless, this study establishes a framework for the analysis of political data in social media in conjunction with German political speeches. The provided corpus can also supplement existing datasets or support new ones incorporating media information. The corpus and software are released under the AGPL license on https://github.com/texttechnologylab/TTLabTweetCorpus.

Ethical Considerations

This study and the developed tweet corpus have been created with attention to ethical considera-

tions. The goal of this work is to provide a large-scale corpus of tweets from German political figures and to make the data collection process as straightforward as possible, while also including media information such as images and videos. Although the corpus is based on publicly available tweets, some individual posts may contain content that is offensive, sensitive, or potentially harmful. Researchers using the corpus should be aware of this when conducting analyses. Similarly, the emotion, sentiment, and topic annotations generated by models may include biases or errors, and their outputs should be interpreted with caution. It is acknowledged that the processing pipeline could theoretically be applied to restricted or prohibited content; however, the intention of this work is to support responsible and transparent scientific research.

Limitations

This work has several limitations. First, the user-based subset is restricted to tweets from German political figures, whereas the search-based subset contains tweets retrieved through hashtag-based queries and is not restricted to these figures. Consequently, neither subset should be interpreted as a complete or fully representative sample of the respective social media discourse. Second, the corpus depends on data that could be collected from X at the time of crawling. As a result, deleted, protected, unavailable, or otherwise inaccessible tweets are not included. The two collection methods also introduce different biases: the user-based method primarily retrieves text-only tweets, whereas the search-based method mainly retrieves tweets containing media. Third, the automatic annotations are produced by existing text-based models and a VLM-based prompting approach. These annotations may contain model-specific errors and biases, especially for politically sensitive, ambiguous, ironic, or multimodal content. The VLM annotations are based on a single prompt, which may limit robustness across different types of images and videos. Finally, the manual evaluation of media-based annotations is small in scale. Because the model outputs were visible while the manual labels were assigned, the annotations may have been influenced by the displayed predictions. The comparison should therefore be interpreted as an exploratory evaluation rather than as a fully independent gold-standard assessment. Although the evaluation indicates that media information can affect human judgments and model performance, the results vary across classification tasks. Further independent manual annotation is therefore required to assess the quality of the automatically generated labels more comprehensively.

Acknowledgments

This work was supported by the research group *Critical Online Reasoning in Higher Education* (CORE, FOR 5404, grant number 462702138) sub-projects B05 (520621868) and C08 (520631675), the Infrastructure Priority Programme *New Data Spaces for the Social Sciences* (SPP 2431, grant number 539634240) sub-project ENTAILab, and the project *Ausbau und Konsolidierung des Fachinformationsdienstes Biodiversitätsforschung* (BIOfid, 326061700) all funded by the German Research Foundation (DFG).

References

Giuseppe Abrami, Mevlüt Bagci, Leon Hammerla, and Alexander Mehler. 2022. German parliamentary corpus (GerParCor). In *Proceedings of the Language Resources and Evaluation Conference*, pages 1900–1906, Marseille, France. European Language Resources Association.

Giuseppe Abrami, Mevlüt Bagci, and Alexander Mehler. 2024. German parliamentary corpus (GerParCor) reloaded. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7707–7716, Torino, Italy. ELRA and ICCL.

Giuseppe Abrami, Markos Genios, Filip Fitzermann, Daniel Baumartz, and Alexander Mehler. 2025. Docker unified UIMA interface: New perspectives for NLP on big data. *SoftwareX*, 29:102033.

Giuseppe Abrami, Manuel Stoeckel, and Alexander Mehler. 2020. TextAnnotator: A UIMA based tool for the simultaneous and collaborative annotation of texts. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 891–900, Marseille, France. European Language Resources Association.

Ali Abusaleh. 2025. Multimodal inference as DUUI component. https://github.com/texttechnologylab/duui-uima/tree/main/duui-mm.

AfD-Fraktion im Deutschen Bundestag. 2026. AfD-fraktion im bundestag. Last accessed: 2026-08-07.

Dimosthenis Antypas, Asahi Ushio, Francesco Barbieri, and Jose Camacho-Collados. 2024. Multilingual topic classification in X: Dataset and analysis. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20136–20152, Miami, Florida, USA. Association for Computational Linguistics.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*.

Ashwin Balasubramanian, Vito Zou, Hitesh Narayana, Christina You, Luca Luceri, and Emilio Ferrara. 2025. A public dataset tracking social media discourse about the 2024 U.S. presidential election on Twitter/X. In *Proceedings of the 6th International Workshop on Cyber Social Threats (CySoc 2025), held in conjunction with the 19th International AAAI Conference on Web and Social Media*. Association for the Advancement of Artificial Intelligence.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

Francesco Barbieri, Jose Camacho-Collados, Francesco Ronzano, Luis Espinosa-Anke, Miguel Ballesteros, Valerio Basile, Viviana Patti, and Horacio Saggion. 2018. SemEval 2018 task 2: Multilingual emoji prediction. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 24–33, New Orleans, Louisiana. Association for Computational Linguistics.

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.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Katja Belousova, Oliver Klein, and Franca Rexeis. 2025. Wahl anfechten? BSW-idee juristisch heikel. ZDFheute. Published on 2025-02-24; Last accessed: 2026-08-07.

Lorenzo Bertolini, Adriana Michalak, and Julie Weeds. 2024. DReAMy: a library for the automatic analysis and annotation of dream reports with multilingual large language models. *Sleep Medicine*, 115:406–407. Abstracts from the 17th World Sleep Congress.

Elizabeth Blakey. 2024. The day data transparency died: How Twitter/X cut off access for social research. *Contexts*, 23(2):30–35.

Laura Ana Maria Bostan, Evgeny Kim, and Roman Klinger. 2020. GoodNewsEveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1554–1566, Marseille, France. European Language Resources Association.

Ruth Breeze. 2020. Exploring populist styles of political discourse in Twitter. *World Englishes*, 39(4):550–567.

BSW-Gruppe im Bundestag. 2026. Bündnis Sahra Wagenknecht: BSW-Gruppe im Bundestag. Last accessed: 2026-08-07.

Daniel Bundan, Giuseppe Abrami, and Alexander Mehler. 2025. Multimodal Docker unified UIMA interface: New horizons for distributed microservice-oriented processing of corpora using UIMA. In *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Long and Short Papers*, KONVENS '25, pages 257–268, Hannover, Germany. HsH Applied Academics.

Bundestagsfraktion Bündnis 90/Die Grünen. 2026. Bündnis 90/Die Grünen Bundestagsfraktion. Last accessed: 2026-08-07.

Jose Camacho-Collados, Kiamehr Rezaee, Talayeh Riahi, Asahi Ushio, Daniel Loureiro, Dimosthenis Antypas, Joanne Boisson, Luis Espinosa Anke, Fangyu Liu, and Eugenio Martínez Cámara. 2022. TweetNLP: Cutting-edge natural language processing for social media. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–49, Abu Dhabi, UAE. Association for Computational Linguistics.

Bruno Castanho Silva and Sven-Oliver Proksch. 2021. Fake it 'til you make it: A natural experiment to identify european politicians' benefit from Twitter bots. *American Political Science Review*, 115(1):316–322.

Bruno Castanho Silva and Sven-Oliver Proksch. 2022. Politicians unleashed? political communication on Twitter and in parliament in western europe. *Political Science Research and Methods*, 10(4):776–792.

Santiago Castro. 2017. Fast Krippendorff: Fast computation of Krippendorff's alpha agreement measure. `https://github.com/pln-fing-udelar/fast-krippendorff`.

CDU/CSU-Fraktion. 2026. CDU/CSU-Fraktion im Deutschen Bundestag. Last accessed: 2026-08-07.

Yiqun T. Chen and James Zou. 2023. TWIGMA: a dataset of AI-generated images with metadata from Twitter. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

CitizenLab (CitizenLabDotCo). 2022. twitter-xlm-roberta-base-sentiment-finetunned. Hugging Face model repository. Last accessed: 2026-08-07.

Michael D. Conover, Bruno Goncalves, Jacob Ratkiewicz, Alessandro Flammini, and Filippo Menczer. 2011. Predicting the political alignment of Twitter users. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 192–199.

Alexander Philip Dawid and Allan M Skene. 1979. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):20–28.

Tom De Smedt and Sylvia Jaki. 2018. The polly corpus: Online political debate in germany. In *Proceedings of the 6th conference on computer-mediated communication (CMC) and social media corpora (CMC-corpora 2018)*, pages 33–36.

Leon Derczynski, Kalina Bontcheva, and Ian Roberts. 2016. Broad Twitter corpus: A diverse named entity recognition resource. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1169–1179, Osaka, Japan. The COLING 2016 Organizing Committee.

Deutscher Bundestag. 2026. Deutscher bundestag parlament fraktionen. Last accessed: 2026-08-07.

Joseph Leon Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Félix-Antoine Fortin, François-Michel De Rainville, Marc-André Gardner Gardner, Marc Parizeau, and Christian Gagné. 2012. DEAP: evolutionary algorithms made easy. *J. Mach. Learn. Res.*, 13(1):2171–2175.

Fraktion der Freien Demokraten im Deutschen Bundestag. 2026. FDP-Fraktion im Bundestag – archiv. Last accessed: 2026-08-07.

Fraktion Die Linke im Bundestag. 2026. Die linke im bundestag. Last accessed: 2026-08-07.

Mai Osama Ghoraba. 2023. Influential spanish politicians' discourse of climate change on Twitter: A corpus-assisted discourse study. *Corpus Pragmatics*, 7(3):181–240.

Simon Gonzalez. 2024. ILiAD: An interactive corpus for linguistic annotated data from Twitter posts. *arXiv preprint arXiv:2407.15374*.

Lara Grimminger and Roman Klinger. 2021. Hate towards the political opponent: A Twitter corpus study of the 2020 US elections on the basis of offensive speech and stance detection. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 171–180, Online. Association for Computational Linguistics.

Kadhim Hayawi, Sakib Shahriar, Mohamed Adel Serhani, Ikbal Taleb, and Sujith Samuel Mathew. 2022. ANTi-Vax: a novel Twitter dataset for COVID-19 vaccine misinformation detection. *Public Health*, 203:23–30.

Nils Constantin Hellwig, Markus Bink, Thomas Schmidt, Jakob Fehle, and Christian Wolff. 2023. Transformer-based analysis of sentiment towards German political parties on Twitter during the 2021 election year. In *Proceedings of the 6th International Conference on Natural Language and Speech Processing (ICNLSP 2023)*, pages 84–98, Online. Association for Computational Linguistics.

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

Muhammad Imran, Prasenjit Mitra, and Carlos Castillo. 2016. Twitter as a lifeline: Human-annotated Twitter corpora for NLP of crisis-related messages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1638–1643, Portorož, Slovenia. European Language Resources Association (ELRA).

Aidos Konyspay, Pakizar Shamoi, Malika Ziyada, and Zhusup Smambayev. 2025. Meme similarity and emotion detection using multimodal analysis. In *2025 International Conference on Activity and Behavior Computing (ABC)*. IEEE.

Klaus Krippendorff. 1970. Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, 30(1):61–70.

Taja Kuzman and Nikola Ljubešić. 2025. LLM teacher-student framework for text classification with no manually annotated data: A case study in IPTC news topic classification. *IEEE Access*, 13:35621–35633.

Johannes Langguth, Daniel Thilo Schroeder, Petra Filkuková, Stefan Brenner, Jesper Phillips, and Konstantin Pogorelov. 2023. COCO: an annotated Twitter dataset of COVID-19 conspiracy theories. *Journal of Computational Social Science*, 6(2):443–484.

Alexander Leonhardt, Giuseppe Abrami, Daniel Baumartz, and Alexander Mehler. 2023. Unlocking the heterogeneous landscape of big data NLP with DUUI.

In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 385–399, Singapore. Association for Computational Linguistics.

Christian E. Lopez and Caleb Gallemore. 2021. An augmented multilingual Twitter dataset for studying the COVID-19 infodemic. *Social Network Analysis and Mining*, 11(1):102.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 4768–4777, Red Hook, NY, USA. Curran Associates Inc.

Richard McCreadie, Ian Soboroff, Jimmy Lin, Craig Macdonald, Iadh Ounis, and Dean McCullough. 2012. On building a reusable Twitter corpus. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 1113–1114, New York, NY, USA. Association for Computing Machinery.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

Martin M. Müller and Marcel Salathé. 2019. Crowdbreaks: Tracking health trends using public social media data and crowdsourcing. *Frontiers in Public Health*, 7.

Usman Naseem, Imran Razzak, Matloob Khushi, Peter W. Eklund, and Jinman Kim. 2021. COVIDSenti: A Large-Scale Benchmark Twitter Data Set for COVID-19 Sentiment Analysis. *IEEE Transactions on Computational Social Systems*, 8(4):1003–1015.

OCLC. 2025. Introduction to the Dewey Decimal Classification. `https://www.oclc.org/content/dam/oclc/dewey/versions/print/intro.pdf`. PDF; last updated 29 Sep 2025. Accessed 11 Oct 2025.

Bolane Olaniran and Indi Williams. 2020. Social media effects: Hijacking democracy and civility in civic engagement. In John Jones and Michael Trice, editors, *Platforms, Protests, and the Challenge of Networked Democracy*, pages 77–94. Palgrave Macmillan.

Jayanta Paul, Siddhartha Mallick, Abhijit Mitra, Anuska Roy, and Jaya Sil. 2025. Multi-modal Twitter data analysis for identifying offensive posts using a deep cross-attention-based transformer framework. *ACM Trans. Knowl. Discov. Data*, 19(3).

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, 12:2825–2830.

Saša Petrović, Miles Osborne, and Victor Lavrenko. 2010. The Edinburgh Twitter corpus. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*, pages 25–26, Los Angeles, California, USA. Association for Computational Linguistics.

Christian Rauh and Jan Schwalbach. 2020. The ParlSpeech V2 data set: Full-text corpora of 6.3 million parliamentary speeches in the key legislative chambers of nine representative democracies.

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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada. Association for Computational Linguistics.

Philipp Schmid. 2022. distilbert-base-multilingual-cased-sentiment. Hugging Face model repository. Last accessed: 2026-08-07.

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 *SciPy (Proceedings of the 9th Python in Science Conference)*.

Juan Carlos Medina Serrano, Morteza Shahrezaye, Orestis Papakyriakopoulos, and Simon Hegelich. 2019. The rise of germany's AfD: A social media analysis. In *Proceedings of the 10th International Conference on Social Media and Society*, SMSociety '19, pages 214–223, New York, NY, USA. Association for Computing Machinery.

Shivshankar (02shanky). 2023. finetuned-twitter-xlm-roberta-base-emotion. Hugging Face model repository. Last accessed: 2026-08-07.

SPD-Bundestagsfraktion. 2026. SPD-Bundestagsfraktion. Last accessed: 2026-08-07.

Gaurish Thakkar, Sherzod Hakimov, and Marko Tadić. 2024. M2SA: Multimodal and multilingual model for sentiment analysis of tweets. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10833–10845, Torino, Italia. ELRA and ICCL.

Tolga Uslu, Alexander Mehler, Andreas Niekler, and Daniel Baumartz. 2018. Towards a DDC-based topic network model of Wikipedia. In *Proceedings of 2nd International Workshop on Modeling, Analysis, and Management of Social Networks and their Applications (SOCNET 2018), February 28, 2018*.

Dmitry Ustalov, Nikita Pavlichenko, and Boris Tseitlin. 2024. Learning from Crowds with Crowd-Kit. *Journal of Open Source Software*, 9(96):6227.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. SemEval-2018 task 3: Irony detection in English tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 39–50, New Orleans, Louisiana. Association for Computational Linguistics.

Maximilian Weissenbacher and Udo Kruschwitz. 2024. Analyzing offensive language and hate speech in political discourse: A case study of German politicians. In *Proceedings of the Fourth Workshop on Threat, Aggression & Cyberbullying @ LREC-COLING-2024*, pages 60–72, Torino, Italia. ELRA and ICCL.

Tobias Widmann and Maximilian Wich. 2023. Creating and comparing dictionary, word embedding, and transformer-based models to measure discrete emotions in german political text. *Political Analysis*, 31(4):626–641.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Jerrold H. Zar. 2014. Spearman rank correlation: Overview. In *Wiley StatsRef: Statistics Reference Online*. John Wiley & Sons, Ltd.

A VLM Prompt

Listing 1 shows the VLM prompt design.

Listing 1: Abstract-level Prompt for Media-based classification

```
You are an AI system that performs
    multimodal understanding of visual
    input.
For a given image/video, predict:
1. Emotion (from a fixed label set)
2. Sentiment (positive, neutral, or
    negative)
3. Topic (using DDC Level 1
    classification)
For each, return:
- predicted label
- confidence distribution over possible
    classes
- explanation grounded in the image
 Ground all outputs in visible content
     only.
```