Model in Distress: Sentiment Analysis on French Synthetic Social Media

Pierre-Carl Langlais[1] **Pavel Chizhov**[1,3] **Yannick Detrois**[1,4] **Carlos Rosas Hinostrosa**[1]
Ivan P. Yamshchikov[1,3] **Bastien Perroy**[2]

[1]PleIAs, Paris, France [2]Passenger Cognition Lab, RATP Group, Paris, France
[3]CAIRO, THWS, Würzburg, Germany [4]EPFL, Lausanne, Switzerland

Correspondence: pierre-carl@pleias.fr

Abstract

Automated analysis of customer feedback on social media is hindered by three challenges: the high cost of annotated training data, the scarcity of evaluation sets, especially in multilingual settings, and privacy concerns that prevent data sharing and reproducibility. We address these issues by developing a generalizable synthetic data generation pipeline applied to a case study on customer distress detection in French public transportation. Our approach utilizes backtranslation with fine-tuned models to generate 1.7 million synthetic tweets from a small seed corpus, complemented by synthetic reasoning traces. We train 600M-parameter reasoners with English and French reasoning that achieve 77-79% accuracy on human-annotated evaluation data, matching or exceeding SOTA proprietary LLMs and specialized encoders. Beyond reducing annotation costs, our pipeline preserves privacy by eliminating the exposure of sensitive user data. Our methodology can be adopted for other use cases and languages.

🤗 DistressedModel/Distressed-Model-Data

1 Introduction

Large companies and services handle a substantial volume of reviews and social media mentions. Social media has become a primary means of filing complaints due to the need for unfiltered, independent feedback and ease of use (SocialMediaToday, 2014; Causon, 2023; Manner and Lane, 2018). This convenience has driven a corresponding increase in the volume of complaints. For example, by 2014, complaints in the United Kingdom had doubled compared to 2013, with 31% expressed through social media (Istanbulluoglu, 2017). Social media also provides anonymity and eliminates direct confrontation, enabling customers to express frustration while exerting perceived public pressure that private reviews may lack (Arora et al., 2025). Furthermore, if customers complain offline or in private, they may then opt for a second round of complaint online (Frasquet et al., 2021).

As customer expectations around social media engagement grow, response time has become critically important (Sigurdsson et al., 2021). For instance, it was recently shown to influence the chances of repurchase (Istanbulluoglu and Sakman, 2024). Therefore, it is crucial for a company to quickly detect the issues and react accordingly by implementing a service recovery (Istanbulluoglu and Oz, 2023). When services operate at scale, and social media interactions exceed manual processing capacity, automated tools such as language models offer a practical solution.

Language models of different sizes and with different design purposes are being developed daily. Proprietary large language models (LLMs) might act as agents for such purposes, while an alternative solution would be specialized small language models. The benefits of LLMs typically include their versatility and ability to understand complex contexts, while specialized small language models are favored for their robustness and, what is especially relevant to public institutions, for the traceability of open-source solutions and an opportunity for autonomous deployment.

A crucial part of automatic review evaluation is sentiment analysis, which typically involves detecting a wide range of emotions and distress to identify the purpose and sentiment of the user messages and classify them into categories required by the business. The models for such a classification task require training on a substantial portion of representative data and reliable annotations, which are often costly to obtain. Furthermore, the trained models need to be properly evaluated, which is often problematic due to the specificity of each case and the lack of comprehensive benchmarks, especially for non-English languages. Even though modern machine translation systems are very advanced, simply translating all the training data into

Findings of the Association for Computational Linguistics: ACL 2026, pages 43009–43021
July 2-7, 2026 ©2026 Association for Computational Linguistics

the required language would degrade data quality (Arnett, 2025), especially given the domain's complex style.

Synthetic data generation has recently emerged as a promising method for training specialized models in regulated sectors with the lack of annotated data (Hughes, 2024; Mendes et al., 2025; Meyer and Corneil, 2025; Tan et al., 2025). Pipelines provide privacy and compliance by design, as personal or sensitive data is never exposed; instead, it is substituted with realistic simulations.

In this work, we introduce one of the first large-scale synthetic pipelines for customer reviews on social media. The pipeline can be applied across various domains and languages to address the aforementioned challenges. We demonstrate its effectiveness on the example of Paris public transport. Our key contributions include:

- We collect a small dataset of tweets annotated by both native speakers and LLMs and aggregate these annotations into reliable labels.
- We present a methodology for creating a synthetic pipeline for review generation using backtranslation and synthetic reasoning, which is applicable to other domains, and release a large collection of synthetic tweets[1].
- We train small reasoning models using synthetic data and evaluate them using human-annotated data, reaching a performance comparable to or better than that of modern LLMs.

2 Related Work

Previous sentiment analysis research was largely concentrated on English data. TweetEval (Barbieri et al., 2020) combined several SemEval tasks into a compound benchmark for sentiment analysis (Rosenthal et al., 2017), emotion recognition (Mohammad et al., 2018), stance (Mohammad et al., 2016), irony (Van Hee et al., 2018), hate speech (Basile et al., 2019), and offensive language detection (Zampieri et al., 2019), as well as emoji prediction (Barbieri et al., 2018). All the subsets were exclusively in English. Despite the initial prevalence of small encoder models applied to this task, it was recently used to evaluate large language models (Arslan, 2025).

Previous efforts for non-English data collection and classification included tweet sentiment analysis in 14 African languages (Muhammad et al., 2023), multi-modal analysis in 21 languages (Thakkar et al., 2024), and other approaches in Eastern European languages (Filip et al., 2025), Bengali (Sazzed, 2020), Indonesian (Geni et al., 2023), and Hausa (Aliyu et al., 2025).

Encoder-style solutions. Typical solutions for the classification tasks of such sort are small BERT-like models (Hoque et al., 2024; Moura et al., 2024), or classical machine learning approaches to building embeddings (TF-IDF, Bag of Words, Word2Vec, etc.) and classification algorithms (Logistic Regression, Random Forest, etc.) (Agrawal et al., 2025), as well as custom pipelines combining both approaches (Mudarakola et al., 2025).

Decoder-style solutions. There has also been a growing line of work where decoder-style language models are used for sentiment analysis and review classification (Suen et al., 2025; Elmitwalli et al., 2024). Even though these approaches are more computationally intensive, which is partially mitigated by recent advancements in small language model development (Gemma Team et al., 2025; Yang et al., 2025), they often provide additional features such as custom prompts (Suen et al., 2025) and reasoning (Ahmed et al., 2026).

3 Task and Motivation

In this work, we analyze public transport reviews on Twitter in French and aim to build a **distress detection** tool that enables support agents to quickly monitor issues expressed on social media. While the task is similar to an ordinary classification pipeline, we approach it in a non-standard way by building a synthetic tweet generation pipeline and training small reasoning models. This is motivated by the following factors:

- **Imbalance:** operational classification of distress is different from the abstract definition and concerns a very small selection of tweets. Using only organic data would put a ceiling on the capabilities of small models.
- **Data drift:** even though a large historical dataset of 30 million tweets collected from 2012 onward (see Appendix A) could be leveraged to mitigate data scarcity, there have been substantial changes in user expression and collection design. Thus, any models trained on these data might perform suboptimally on more recent data.

[1]We release the anonymized synthetic tweets on HuggingFace: 🤗 DistressedModel/Distressed-Model-Data.

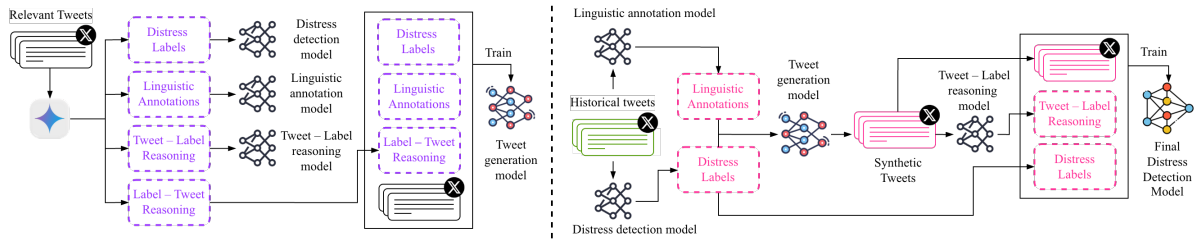

Figure 1: Synthetic pipeline and distress detection model training. **Left:** auxiliary model fine-tuning. **Right:** synthetic pipeline for generating synthetic tweets and training the distress detection model. All LLM-generated data are shown in purple. All pipeline-generated synthetic data are shown in pink.

- **Data sensitivity:** tweets have always been a gray area regarding GDPR, as, despite being public expressions, they might contain identifiable or sensitive information on private persons. Moreover, monitoring also typically includes strictly private user data (application feedback, phone calls) that could never be disclosed for training.
- **Generalization:** The monitoring infrastructure anticipates a range of extreme events that never occurred in the historical data, which can be handled by language models with generalist knowledge and reasoning capabilities.

Given the unique challenge of realistic, primarily non-English tweet generation, we build our synthetic pipeline around custom finetuned models. Thus, we approach this problem via **specialist distillation** (Liu, 2025), when a specialized model is trained to produce realistic synthetic data, which is then used to train the final model.

4 Methodology

Data. The original dataset comprises more than 30 million tweets related to French public transport between 2012 and 2025, from over 5 million users. Data acquisition during this period proceeded in separate phases. During the **first phase** (2012 to June 2023), a Twitter research license permitted large-scale data scraping of tweets that mention French public transport or are replies to threads referencing it. This ensures exhaustive coverage but also produces noisy data, including possibly unrelated tweets. This phase accounts for the largest part of our data, with peak monthly tweet counts exceeding 1.5 million. The **second phase** (November 2024 and onward) followed Twitter's acquisition and evolution of social media infrastructure. A professional API license was acquired to collect the most recent and relevant tweets in smaller volumes (on average 30 000 tweets per month). While the majority of tweets are in French (66.3%), English is also represented (15.3%), with other less prominent languages making up the remaining fraction. The complete details about the dataset are presented in Appendix A.

Synthetic pipeline. The core of our synthetic pipeline (Figure 1) relies on **backtranslation**. We use Gemini 3 Pro (Doshi, 2025) as a convenience to generate our 3 000 seed annotations (see prompts in Appendix E) using a sample from the second phase of our dataset. We use a proprietary model for this generation, although recent capable open-weight models can serve as an effective substitute (Sutro, 2026; Emberson, 2025). At the initial stage of the pipeline (left part of Figure 1), we annotate a set of tweets with distress labels and a set of features, combining linguistic characteristics, emotional valence, and general contextual information (event, location, and any identification of the author). We also generate synthetic reasoning traces that lead to tweets from their annotations, and vice versa. We then use these data to train specialized annotation and reasoning models that will be used as parts of the synthetic pipeline. For reasoning traces, we use both English and French in separate pipelines, yielding two different reasoner models. To stimulate diversity for the linguistic annotation model, we also add a portion of tweets unrelated to distress. We also train the tweet generation model on a backtranslation-style combination of annotations, reasoning traces, and original tweets. All four auxiliary models are fine-tuned versions of Gemma 3 12B Instruct (Gemma Team et al., 2025).

After this, we use a filtered collection of first-phase tweets as seed to generate a dataset of 1 737 797 synthetic tweets using the fine-tuned models (right part of Figure 1). We separately validate the variety of generated tweets due to the risk

Model	Parameters	Fine-tuning	Accuracy	Precision	Recall	F1
Claude Sonnet 4.5	—	—	72.2	75.8	62.8	68.7
Gemini 2.5 Pro	—	—	71.5	71.1	69.5	70.3
Mistral 3 Large	675B	—	78.5	88.6	64.1	74.4
Qwen3.5	397B	—	67.5	**91.3**	36.8	52.5
Distress Reasoner (ours)	600M	Synth. tweets, EN reasoning	78.0	80.7	71.9	76.0
Distress Reasoner (ours)	600M	Synth. tweets, FR reasoning	**79.5**	82.2	**73.7**	**77.7**
ModernBERT Large	395M	Synth. tweets	79.1	81.0	**74.4**	77.6
CamemBERTv2 Base	110M	Synth. tweets	**79.3**	**81.6**	74.3	**77.7**

Table 1: Accuracy (%), Precision (%), Recall (%), and F1 evaluations for distress detection. We compare LLMs (top rows), our decoder models (middle rows), and fine-tuned encoders (bottom rows). Our decoder models are trained separately with English (EN) and French (FR) reasoning. The best scores among decoders and encoders are **in bold**.

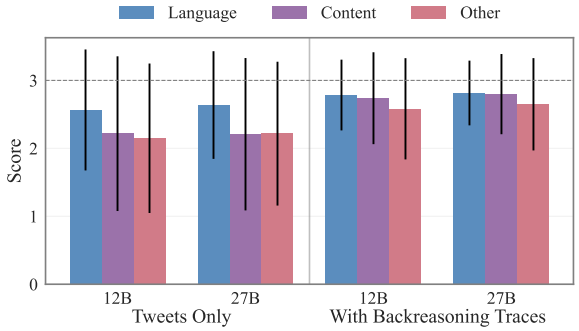

Figure 2: Evaluating Gemma models fine-tuned for tweet generation. Bars represent standard deviation.

that the generation model may reproduce the seed tweets. Of the generated tweets, 14 429 (0.83%) were close to the original tweets, as evaluated with `SequenceMatcher` from `difflib` using a similarity threshold of 65%. We filter these entries out before the training of the final model. We combine the resulting 1 723 368 synthetic tweets with generated reasoning traces and SYNTH[2], an open-source pre-training dataset for generalist reasoning, into a training dataset for our final model, a decoder model with Llama-style architecture and 600 million parameters (see architecture details in Appendix C). The training dataset results in 1 010 748 472 tokens, over which we train for three epochs on four NVIDIA H100 GPUs with an effective batch size of 262 144 tokens: context length of 2 048, per-GPU-batches of size 8, and 8 gradient accumulation steps.

Evaluation sample. From the second-phase subsample of the dataset, we select 2 000 relevant tweets using annotations by three LLMs. The selected tweets do not intersect with the fine-tuning sample in the synthetic pipeline. We then slice the selected sample into 40 batches of 50 tweets and hand them to a group of annotators, so that each tweet receives from 9 to 11 binary distress annotations, resulting in an average of 10.2 annotators per tweet (see Appendix B for the complete annotation details). We aggregate the labels through majority voting, resolving ties in favor of the distress label. We leverage this sample for the evaluation of the final distress detection model.

[2] https://huggingface.co/datasets/PleIAs/SYNTH

5 Experimental Results

Auxiliary model ablation. We validate the tweet annotation model using Gemini 3 Pro as a judge to assess how a generated tweet aligns with the generation instructions (see evaluation instructions in Appendix D). We present a comparison of two Gemma versions (12B and 27B), fine-tuned with and without synthetic reasoning traces, in Figure 2. The results show that synthetic reasoning traces improve the correspondence to factual labels in the generation instructions, while language quality remains high in both modes. The difference between the 12B and 27B models is negligible; therefore, we utilize the 12B version for speed.

Distress detection. We present the results of evaluating on the human-annotated sample of tweets in Table 1. We compare our small models to proprietary LLMs: Claude Sonnet 4.5 (Anthropic, 2025) and Gemini 2.5 Pro (Comanici et al., 2025), and large open-weight LLMs: Mistral 3 Large (Mistral AI, 2025) and Qwen3.5 397B (Qwen Team, 2026). We evaluate our models with low temperature and repetition penalty values (both set to 0.1) and treat parsing failures as negative predictions. Despite their small size, the newly trained specialized models achieve competitive performance on the constructed task, comparable to or better than heavy multilingual LLMs with reasoning capabil-

ities. Mistral 3 Large achieves the closest performance, possibly due to better adaptation to real-world French data, yet has inferior recall, which is crucial in complaint-handling scenarios. Qwen3.5 has the highest precision, but the lowest recall. We hypothesize that, although the model has multilingual capabilities, it is not well-adapted to complicated non-English Twitter slang.

We separately fine-tune two encoder models on our synthetic training set: ModernBERT Large (Warner et al., 2025) and CamemBERTv2 Base (Antoun et al., 2024)[3]. Both models performed comparably to our best decoder model, with CamemBERT performing slightly better, likely due to the better adaptation to French content. This shows that our synthetic pipeline is also suitable for training specialized encoders; however, encoder models lack reasoning traces, which are highly practical for real-world scenarios. We further discuss this in Section 6.

6 Discussion

Our results demonstrate that a small set of relevant real-world data can be used as a seed for the synthetic generation of reliable data, including synthetic reasoning. This approach made it possible to share the training data, ensuring better standards of reproducibility and fostering similar work on sensitive data resources or annotation. The data generated at scale exhibits diversity and can then serve as a training set for a classifier that will be suitable for training a real-world distress detection model that achieves state-of-the-art performance comparable to or better than that of proprietary LLMs. The demonstrated methods can be used to scale up the training data in conditions of low availability of annotated real-world data. The resulting model can then be autonomously deployed without reliance on a proprietary commercial API and efficiently process large portions of reviews due to its small size (600 million parameters).

The comparable performance of encoder-based models demonstrates that the pipeline can be used for encoder-based scenarios as well. However, a decoder-based reasoner offers two advantages over a scalar encoder. First, conditioning the label on an explicit reasoning trace improves correspondence with the fine-grained content annotations requested at generation (Figure 2). Second, the reasoning traces persist during model inference as usable outputs for agents acting on each decision. This aligns with the argumentative view of reasoning, on which reasoning is fundamentally an artifact produced to be evaluated by others (Mercier and Sperber, 2011): an opaque score offers nothing to scrutinize, whereas a trace exposes the inferences a downstream agent can accept, contest, or override. In a cognitive task analysis of three rail control rooms, Dadashi et al. (2021) make the parallel operational point that automation should support exploration of the reasoning behind its decisions rather than expose only an alert. The same trace also externalizes which patterns the model has learned to associate with distress, complementing the privacy-preserving design of Section 3 with the form of transparency that European frameworks such as the GDPR and the AI Act increasingly require of automated decision systems acting on personal data. In the public transport setting, the actions taken on the basis of a flagged tweet can range from a routine acknowledgment to consequential interventions such as dispatching emergency staff or evacuating passengers from a train stalled in a tunnel; in this regime, the trace lets a support agent inspect the cited spans rather than accept or reject an opaque score. This is an affordance no BERT-style classifier provides.

7 Conclusion

Our work provides a proof of concept that, under a specialist synthetic pipeline, a small language model with reasoning can achieve performance comparable to or better than state-of-the-art large models, both proprietary and open-weight. Such a model is beneficial not only from the perspective of specialization and task-tailoring, but also enables fast processing of large volumes of requests and fully autonomous deployment. The decoder design with reasoning traces also offers practical benefits over encoder-based models. The demonstrated pipeline is relatively universal and can be adapted to various domains and languages, given a small set of seed data and a capable auxiliary language model that can generate the seed synthetic data in the required language.

Limitations

Our work demonstrates results only on one classification characteristic (distress) and only in a specific domain and language (public transport reviews in

[3]We release the fine-tuned encoder-style models as 🤗 `DistressedModel/modernbert-distress-signal` and 🤗 `DistressedModel/camembert-distress-signal`

French). Thus, the model's application to other tasks and domains would require separate fine-tuning and additional synthetic data generation.

The demonstrated approach requires a certain amount of real-world data as a seed for fine-tuning auxiliary models and generating synthetic data. Thus, the pipeline is not suitable for cases where real-world data is unavailable. Furthermore, since our original annotations and synthetic reasoning were generated by Gemini, this approach may not be suitable for private real-world data. In such cases, practitioners would need a locally deployed LLM for auxiliary data generation; however, the results from such a setup might be just as good given the right model.

We used Gemma as a base model for synthetic fine-tuning, which is essentially strong in terms of multilinguality. However, if a new pipeline requires expertise in an extremely low-resource language, a suitable fine-tuning model should be selected. This is especially important when, as in our case, the task requires strong generalizability to the unseen cases and understanding of complicated language nuances and slang.

Acknowledgements

The computational resources for this work were provided by Jean Zay supercomputer (project EuroSynthPlay, compute grant AD011016886R1).

References

Renuka Agrawal, Mehuli Majumder, Ishita Yadav, Nandini Taneja, Safa Hamdare, and Preeti Hemnani. 2025. Evaluating sentiment analysis models: A comparative analysis of vaccination tweets during the COVID-19 phase leveraging DistilBERT for enhanced insights. *MethodsX*, 14:103407.

Kanwal Ahmed, Muhammad Imran Nadeem, Guanghui Wang, Fang Zuo, and Zhijie Han. 2026. LLM-infused multi-module transformer for emotion-aware sentiment analysis in few-shot scenarios. *Information Fusion*, 126:103668.

Yusuf Aliyu, Aliza Sarlan, Kamaluddeen Usman Danyaro, Abdullahi Sani abd Rahman, Aminu Aminu Muazu, and Mustapha Yusuf Abubakar. 2025. Deep learning techniques for sentiment analysis in code-switched Hausa-English tweets. *International Journal of Information Management Data Insights*, 5(1):100330.

Anthropic. 2025. Introducing Claude Sonnet 4.5. https://www.anthropic.com/news/claude-sonnet-4-5. Accessed: 2026-04-17.

Wissam Antoun, Francis Kulumba, Rian Touchent, Éric de la Clergerie, Benoît Sagot, and Djamé Seddah. 2024. CamemBERT 2.0: A Smarter French Language Model Aged to Perfection. *Preprint*, arXiv:2411.08868.

Catherine Arnett. 2025. Best Practices for Open Multilingual LLM Evaluation.

Swapan Deep Arora, Anirban Chakraborty, and Gopalakrishnan Narayanamurthy. 2025. Why and When Consumers Post Complaint Messages on Social Media? Conceptualizing Social Voice as a Distinct Complaining Behaviour. *British Journal of Management*, 36(4):1746–1766.

Ezgi Arslan. 2025. Sentiment Analysis Benchmark Testing: ChatGPT, Claude & DeepSeek. Accessed: 2025-12-02.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa-Anke, and Leonardo Neves. 2020. TweetEval:Unified Benchmark and Comparative Evaluation for Tweet Classification. In *Proceedings of Findings of EMNLP*.

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.

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.

Jo Causon. 2023. Social media as a review channel. Accessed: 2025-12-02.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *Preprint*, arXiv:2507.06261.

Nastaran Dadashi, David Golightly, and Sarah Sharples. 2021. Modelling decision-making within rail maintenance control rooms. *Cogn. Technol. Work*, 23(2):255–271.

Rohan Doshi. 2025. Gemini 3 Pro: the frontier of vision AI. Accessed: 2025-12-30.

Sherif Elmitwalli, John Mehegan, Allen Gallagher, and Raouf Alebshehy. 2024. Enhancing sentiment and intent analysis in public health via fine-tuned large language models on tobacco and e-cigarette-related tweets. *Frontiers in Big Data*, 7:1501154.

Luke Emberson. 2025. Open-weight models lag state-of-the-art by around 3 months. https://epoch.ai/data-insights/open-weights-vs-closed-weights-models. Accessed: April 16, 2026.

Tomáš Filip, Martin Pavlíček, and Petr Sosik. 2025. Tuning of language models in Eastern European languages on Twitter/X.

Marta Frasquet, Marco Ieva, and Cristina Ziliani. 2021. Complaint behaviour in multichannel retailing: a cross-stage approach. *International Journal of Retail & Distribution Management*, 49(12):1640–1659.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 Technical Report. *Preprint*, arXiv:2503.19786.

Lenggo Geni, Evi Yulianti, and Dana Indra Sensuse. 2023. Sentiment Analysis of Tweets Before the 2024 Elections in Indonesia Using BERT Language Models. *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI)*, 9(3):746–757.

Md. Nesarul Hoque, Umme Salma, Md. Jamal Uddin, Md. Martuza Ahamad, and Sakifa Aktar. 2024. Exploring transformer models in the sentiment analysis task for the under-resource Bengali language. *Natural Language Processing Journal*, 8:100091.

Jason L. Huang, Paul G. Curran, Jessica Keeney, Elizabeth M. Poposki, and Richard P. DeShon. 2012. Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, 27(1):99–114.

Alyssa Hughes. 2024. Improving synthetic data without compromising privacy protection. Accessed: 2026-01-04.

Doga Istanbulluoglu. 2017. Complaint handling on social media: The impact of multiple response times on consumer satisfaction. *Computers in Human Behavior*, 74:72–82.

Doga Istanbulluoglu and Seda Oz. 2023. Service Recovery via Twitter: An Exploration of Responses to Consumer Complaints. *Accounting Perspectives*, 22(4):435–460.

Doga Istanbulluoglu and Ezgi Sakman. 2024. Successful complaint handling on social media predicts increased repurchase intention: The roles of trust in company and propensity to trust. *European Management Journal*, 42(1):11–22.

Aixin et al. Liu. 2025. DeepSeek-V3.2: Pushing the Frontier of Open Large Language Models. *arXiv preprint*. ArXiv:2512.02556 [cs].

Chris Manner and Wilburn Lane. 2018. Who posts online customer reviews? The role of sociodemographics and personality traits. 30:19–32.

Adam W. Meade and S. Bartholomew Craig. 2012. Identifying careless responses in survey data. *Psychological Methods*, 17(3):437–455.

Jorge M. Mendes, Aziz Barbar, and Marwa Refaie. 2025. Synthetic data generation: a privacy-preserving approach to accelerate rare disease research. *Frontiers in Digital Health*, 7. Publisher: Frontiers.

Hugo Mercier and Dan Sperber. 2011. Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, 34(2):57–74.

Yev Meyer and Dane Corneil. 2025. Nemotron-Personas: Improve AI Training With the First Synthetic Personas Dataset Aligned to Real-World Distributions.

Mistral AI. 2025. Introducing Mistral 3. https://mistral.ai/news/mistral-3. Accessed: 2026-04-17.

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.

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.

Ricardo Moura, Jonnathan Carvalho, Alexandre Plastino, and Aline Paes. 2024. Less is more: Pruning BERTweet architecture in Twitter sentiment analysis. *Information Processing & Management*, 61(4):103688.

Lakshmi Prasad Mudarakola, Ranjith Kumar Gatla, Akella S. Narasimha Raju, Amar Y. Jaffar, Abdullah Alzahrani, and Adis Abebaw Dessalegn. 2025. Multi stage sentiment analysis for product reviews on Twitter using optimized machine learning algorithm. *Scientific Reports*, 15(1):39777.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, Nedjma Ousidhoum, David Ifeoluwa Adelani, Seid Muhie Yimam, Ibrahim Sa'id Ahmad, Meriem Beloucif, Saif M. Mohammad, Sebastian Ruder, Oumaima Hourrane, Pavel Brazdil, Alipio Jorge, Felermino Dário Mário António Ali, Davis David, Salomey Osei, Bello Shehu Bello, Falalu Ibrahim, Tajuddeen Gwadabe, and 8 others. 2023. AfriSenti: A Twitter Sentiment Analysis Benchmark for African Languages. In *Proceedings of the 2023 Conference on Empirical Methods in*

Natural Language Processing, pages 13968–13981, Singapore. Association for Computational Linguistics.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents. https://qwen.ai/blog?id=qwen3.5. Accessed: 2026-04-17.

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.

Salim Sazzed. 2020. Cross-lingual sentiment classification in low-resource Bengali language. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 50–60, Online. Association for Computational Linguistics.

Valdimar Sigurdsson, Nils Magne Larsen, Hulda Karen Gudmundsdottir, Mohammed Hussen Alemu, R. G. Vishnu Menon, and Asle Fagerstrøm. 2021. Social media: Where customers air their troubles—How to respond to them? *Journal of Innovation & Knowledge*, 6(4):257–267.

SocialMediaToday. 2014. The Impact of Online Reviews and Your Business: Positive Only vs. Responding to Negative Reviews. Accessed: 2025-12-02.

Sidney Suen, Ranjan Satapathy, Kenneth Kwok, and Erik Cambria. 2025. Multi-Layered Prompting of Small Language Models for Twitter Sentiment Analysis. Accessed: 2025-12-15.

Team Sutro. 2026. The Future (and Present) of AI is Synthetic Data. Accessed: April 16, 2026.

Bowen Tan, Zheng Xu, Eric P. Xing, Zhiting Hu, and Shanshan Wu. 2025. Synthesizing Privacy-Preserving Text Data via Finetuning without Finetuning Billion-Scale LLMs. In *Forty-second International Conference on Machine Learning*.

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.

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.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 Technical Report. *Preprint*, arXiv:2505.09388.

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.

A Data

In the first collection period between 2012 and 2023, the volume of tweets increased steadily over the years despite an invariable collection scheme, reflecting the growing importance of transport on social media (see Figure 3). With the 2023 API changes, the research license was switched to a basic license with more restrictive collection criteria. Although the new scheme produces a smaller dataset (about 30 000 tweets compared with nearly one million by the end of Period 1), the remaining volume is still sufficient to capture the most important information and identify distress signals. Data from the period between the two phases was not used in this work.

Approximately two-thirds of the dataset consists of French tweets, with the remainder comprising multiple languages, which supports multilingual capabilities (Figure 4). The second-largest language is English, followed by Spanish, with other languages grouped under *Other*, each representing less than 2% of the data. Certain tweets that contain only, or primarily, emojis, media, mentions, hashtags, or links have an undefined language attribute (6.2%). Since 2022, new tags have been introduced to more precisely categorize undefined language attributes. For example, the *Media Links* tag for tweets containing only media links, which account for an additional 2.5% of the data.

B Tweet Annotations

We curated a challenging annotation dataset of real 2 000 tweets. Given the rarity of distress signals, we used an initial selection using the annotations

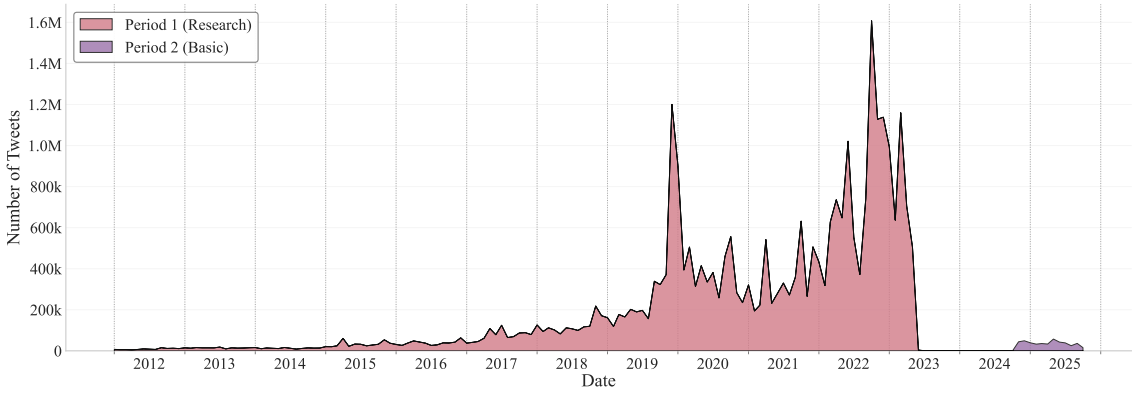

Figure 3: Monthly tweet volume by collection period between 2012 and 2025.

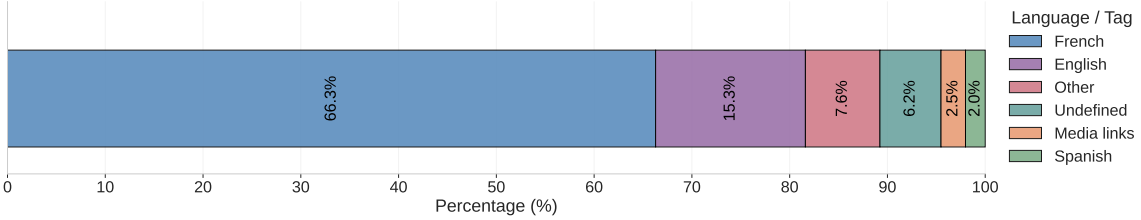

Figure 4: Tweet language distribution in the historical period (Period 1).

from three LLMs. We include the full evaluation set with dehydrated Twitter IDs[4].

We construct our ground truth evaluation dataset using a crowdsourcing strategy, recruiting a cohort of annotators to label the data. The objective is to determine whether a message expressed distress and to quantify the extent of experiential cues provided by the passenger. A total of 2 000 tweets are distributed into batches. Each participant is assigned a batch of 50 tweets and is required to provide a binary classification of distress (Yes/No) and an assessment on a Likert scale of the degree of experiential detail. To mitigate order effects, participants first complete the distress classification for their assigned batch in a randomized order, followed by the experiential cues assessment for the same batch in a newly randomized order.

To ensure data quality, we implement two standard procedures to detect Insufficient Effort Responding (IER). First, following the recommendations of Meade and Craig (2012) for multiple attention check designs, we embed four attention check items (e.g., "Please select 'Yes' for this item") within the survey. Participants who fail more than one check (i.e., fewer than 3 correct answers) are excluded; this criterion led to the exclusion of 4 participants from the initial 415. Second, we screen for rapid responding using the method described by Huang et al. (2012). We calculate the median response time for each participant across all 100 item-level variables. Respondents with a median response time below 2 seconds are flagged as "speeders," as this duration is considered the cognitive minimum required to process a survey item meaningfully. This median-based approach is robust to occasional outliers while effectively identifying consistent careless patterns. This threshold eliminated four additional participants (1.4%).

The final analytical sample consists of 407 participants recruited via convenience sampling. We employ a consensus strategy, using panels of nine to eleven distinct evaluators per message, specifically aiming for a diverse cross-section of usage intensity. The cohort is split between regular riders ($n = 196$, 48.2%) and occasional riders ($n = 211$, 51.8%). We define regular riders as those traveling daily or almost daily ($n = 88$), several times a week ($n = 63$), or several times a month ($n = 45$). Conversely, the occasional group comprises those who travel several times a year ($n = 121$) or less frequently ($n = 90$). This heterogeneity is strategic. While annotators may lack formal expertise in text analysis, their personal history of navigating the network grants them a critical "situated expertise." This personal experience with the transport

[4] 🤗 DistressedModel/Distressed-Model-Data

Category	Setting
Model Architecture	
Number of layers	80
Hidden size	800
Intermediate (FFN) size	2048
Attention heads	10
Key–value heads	5
Activation function	SiLU
Max sequence length	2048
Positional encoding	RoPE ($\theta = 10{,}000$)
Normalization	RMSNorm ($\epsilon = 1 \times 10^{-5}$)
Embedding tying	Enabled
Vocabulary size	65,536
Precision	`bfloat16`
Training Setup	
Optimizer	AdamW
Adam β_1, β_2	0.9, 0.95
Weight decay	0.01
Gradient clipping	0.5
Learning rate	5×10^{-4}
Warmup style	Linear
Decay style	Linear
Warmup steps	500
Decay start step	3000
Minimum learning rate	6×10^{-6}
Training steps	4000
Micro-batch size	4
Gradient accumulation	8
Effective sequence length	2048
Parallelism	
Data parallelism (DP)	4
Tensor parallelism (TP)	1
Pipeline parallelism (PP)	1
ZeRO stage	0

Table 2: Model architecture and training configuration.

environment enables them to intuit the complex, often implicit experiential cues.

Overall agreement among annotators is 80.50% (± 16.70%), which is within expectations given the focus of evaluation on complex cases.

C Architecture and Training

In Table 2, we present the main model architecture and training hyperparameters.

D LLM-as-a-Judge

For the evaluation of how generated tweets correspond to the generation instructions, we use the following rubrics for Gemini 3 Pro (Doshi, 2025), which is used as a judge:

- **Language**: overall linguistic quality, register, code-switching if applicable, and formality.
- **Content**: topic and thematic accuracy, the described issue type, the mentioned audience or target.
- **Other attributes**: emotions, tone, irony, urgency, stance, intent, intensity, and any other stylistic or pragmatic features.

Each criterion is scored on a scale from 0 to 3 with labels:

- 3 = good compliance (expected correct behavior)
- 2 = minor deviations
- 1 = major deviations
- 0 = complete failure or irrelevance

E Generation Prompts

We present prompts used for the generation of the fine-tuning data with an LLM (Gemini 3 Pro; Doshi, 2025) in Figures 5, 6, and 7.

F LLM Usage Statement

In this work, we used AI tools for paraphrasing and grammar correction (Grammarly) and for data visualization design (Claude).

```
You are a **clinician-like annotator** assessing a single tweet for potential **distress and psychometric states**.
You work **bilingually (French/English)**: interpret and quote French text accurately, but write all analysis in English.
Use concise, evidence-based prose.

You have been submitted the following tweet:
{tweet}

You also no in advance the final judgement:
{judgement}

Now try to write a reasoning workflow according to the guidelines below that would converge to this final judgement.

### 1) Pragmatics
* Contextualize: is it a
* Identify tone and stance: literal report / complaint / rhetorical slogan / meme or sarcasm / hyperbole / bureaucratic or formal.
* Who speaks about whom? (self vs. other vs. unspecified).
* Temporal index: now / recent / past / unknown.
* Quote 1–3 exact spans that anchor these judgments.

### 2) Distress

**Goal:** Determine whether any form of distress is present.

* Decide: **present / absent / external**

**Criteria for present:**
* Author is a victim or a direct witness to physical violence or danger. This could be caused by other travellers or agents.
Common issues include being thrown on the rails, attacked by a third party, etc.
* Author is a victim or a direct witness of acts sexual violence, not only assault but also harassment, exhbitionnism.
* Author is a victim or a direct witness health issues, including faintness ("malaise"), breathing (especially is ashtma),
overheating, body compression in highly crowded space.
* Graphic description of violence likely witnessed or even endured by the author (blood, members sectioned, etc.)
* Strong potential for violence or whole situation degenerating. Typically someone having a very disturbing behavior (under drugs, displaying
mental issues, claiming some terrorist intent, being armed, laying something looking like a bomb)
* Immediate and unusual infrastructure risks, even if nothing happened yet: unsecurized rails, flooding, ceiling falling, invasive species
("punaises"), etc.
* Collective distress situation including heat, crowding, discomfort. This may spill into potential or actual fainting, especially involving
fragile people (old, etc.) or collective panic/psychosis.
* Author being a direct witness to train evacuation, (having to break windows, walk on tunnels) or train getting stuck for a long time
(up to hours) with indirect health risks.
* Author showing signs of breakdown, even if the cause might not be dramatic in itself: crying/screaming, disorientation. This can be
expanded by somatic, cognitive, or fragmentation cues.
* Generic yet genuine cry for help, even if details are unclear. So long as the author seems serious and currently enduring this situation

**Criteria for absent:**
* Political rhetoric or commentary with no self-referential distress.
* Light inconvenience expressed through meme, hyperbole, or sarcasm.
* References to self-harm or even suicide ideation that are clearly emphatic.
* Service disruption without spillover risk. This include being stuck in a tunnel for a shortwhile, losing a connection, etc.
* Generic complaint about degraded environment ("racailles").
* Potentially accute annoyance (noise, smells) but without traumatic consequences
* Losing personal objects (bag, wallet, computer) in the subway, unless this translate to accute personal trauma

**Criteria for external:**
* Reference to a clear distress situation but **¨reported** through external sources.
* Downstream consequence of an actual distress situation ("malaise voyageur" causing delays but not happening nearby)
* Any past and resolved event (especially if the person was saved in the end)
* Publication by a media/journalist or other third party.
* Video coverage

Provide a **one-sentence justification** including **verbatim span(s)**.

### 3) Semiotic potential
As a last test, you have to evaluate whether the tweet alone contain enough semiotic signals for a subsequent psychometric assessment.

The cues can be of different nature:
*Explicit relation of a traumatic event.
*Description of inner or supposed mental state.
*Description of psychosomatic symptoms.
*Traumatic connotations in the text itself. This obviously include high emotional valence but even signs of dissociation (like the person
was agressed and write in a super cold tone)

Here your final judgment is a grade, from 1-10.
* 1–3 = figurative or faint
* 4–6 = partial or indirect cues
* 7–10 = explicit, concrete, well-anchored evidence

Produce all results in this order. Never mention that you know the end result in advance, just try to converge to it.

### Text stance & pragmatics ###
[Initial analysis/exploration]

### Distress ###
[Is this relevant for distress? Why? Quote spans. Final answer either **present**, **absent** or **external**]

### Semiotic potential ###
[Primary class + micro-justification with quotes.]
```

Figure 5: Prompt for distress detection and reasoning.

```
I would like to annotate the following tweet:
{tweet}

The annotation is strictly focused on the language analysis. It used the following features.

<event>
A description of the event related in the tweet in English. Should be very brief and hold in one sentence.

<location>
If indicated, name of the location(s) where the event take place. Could be typically subway stations, bus stops, streets, etc.

<time>
If indicated/specified, time of the day or the year the event take place. You only need to be as precise as is mentioned in the text:
could be just hour (9 am) or a complete timestamp.

<persona>
If specified, a short description of what the author of the tweet could be based on text clues. Typically, a middle-aged business person
from Korea, a young French muslim woman, etc. Leave blank if you don't know anything more than a frequent commuter in Paris.

<language_register>
Standardized values can be either:
formal: official complaints, institutional tone, requests for attestations
standard: standardized conversational tone in French, neutral tone, proper grammar,
casual: colloquial form of French for online discussions ("mdr", "vzy", contractions like "j'suis", "y'a")
street: argot, verlan like "ouf", "chelou"
dialectical: regional variety of French, not necessarily in France (also include pidgin etc.)

<language_quality>
standard: generally correct French
imperfect: a few typo, still acceptable for conversational style
poor: hard to parse, showing visible struggle with written French

<language_switching>
If valid, simply list the languages that used in lowercase. If French and English: french, english etc.

<emotion>
The primary emotion conveyed by the tweet. We use the canonical list of emotions from Paul Ekman:
anger, disgust, fear, happiness, sadness, surprise

<tone>
sarcastic: really any form or irony.
emphatic: intense, highly emotional speech.
vulgar: very familiar speech, potentially some insults

<intensity>
The intensity of the complaint:
low: mild complaint
medium: clear frustration
high: strong language, highly emotional
extreme: genuine distress signals

<intent>
service request, complaint, alert, praise, humor, solidarity, practical info

Please only return the annotations field that are valid for this tweet using xml-like tags.

For instance with a short tweet about a health event on line 1 it would be structured like this:
<event>Someone has a heart attack in the station</event>
<location>Saint-Paul station</location>
<time>early morning</time>
<persona>A middle-aged woman from Provence</persona>
<language_register>casual</language_register>
<language_quality>standard</language_quality>
<emotion>fear</emotion>
<intensity>high</intensity>
<theme>health incident</theme>

Instead with a sex agression, we could have something like this:
<event>The author of the tweet is victim of a sexual agression</event>
<location>Garge de Cergy-Pontoise</location>
<time>November 9th</time>
<persona>A red-headed woman</persona>
<language_register>formal</language_register>
<language_quality>standard</language_quality>
<language_switching>french, english</language_switching>
<emotion>sadness</emotion>
<intensity>high</intensity>
<intent>alert</intent>
<theme>sexual violence</theme>

### Analysis ###
[Short analysis/recontextualization of the tweet that should ultimately converge on the features]

### Annotation ###
[The final annotation using the standard fields without anything else. Especially, DO NOT include the text]
```

Figure 6: Prompt for linguistic annotation.

```
You have been submitted this tweet:
{text}

This tweet has been annotated like this:
{annotation}

The annotation follows on a specific set of guidelines.

<event>
A description of the event related in the tweet in English.

<location>
If indicated, name of the location(s) where the event take place.

<time>
If indicated/specified, time of the day or the year the event take place.

<persona>
If specified, a short description of what the author of the tweet could be based on text clues.

<language_register>
Standardized values can be either:
formal: official complaints, institutional tone, requests for attestations
standard: standardized conversational tone in French, neutral tone, proper grammar,
casual: colloquial form of French for online discussions ("mdr", "vzy", contractions like "j'suis", "y'a")
street: argot, verlan like "ouf", "chelou"
dialectical: regional variety of French, not necessarily in France (also include pidgin etc.)

<language_quality>
standard: generally correct French
imperfect: a few typo, still acceptable for conversational style
poor: hard to parse, showing visible struggle with written French

<language_switching>
If valid, simply list the languages that used in lowercase. If French and English: french, english etc.

<emotion>
The primary emotion conveyed by the tweet. We use the canonical list of emotions from Paul Ekman:
anger, disgust, fear, happiness, sadness, surprise

<tone>
sarcastic: really any form or irony.
emphatic: intense, highly emotional speech.
vulgar: very familiar speech, potentially some insults

<intensity>
The intensity of the complaint:
low: mild complaint
medium: clear frustration
high: strong language, highly emotional
extreme: genuine distress signals

<intent>
service request, complaint, alert, praise, humor, solidarity, practical info

Finally you also have a specific entry distress:

<distress>
absent
present

**Criteria for present:**
<...>

**Criteria for absent:**
<...>

Now you have to create a draft leading to the annotation, from the tweets. Even though you already have the final annotations,
it should really be as if you are just focused on extrapolating them from the tweet.

The draft should avoid excessive bold and bullet point, be rather structured like a well stylized text flow.
It should condensed and to the point, not overdo it.

Before starting the draft, you should briefly assess the task at hand, based on the annotations to predict and the text as input.
You should especially estimate how long the draft should be.

You should format your answer like this:

<analysis>
[short analysis of the way you would structure the draft and assess its length given the complexity of the tweet.]
</analysis>

<draft>
[your draft leading to the tweet]
</draft>
```

Figure 7: Draft generation prompt. Criteria for present/absent are similar to those in Figure 5.