A Retail-Corpus for Aspect-Based Sentiment Analysis with Large Language Models

Oleg Şilcenco[1], Marcos R. Machado[1], Wallace C. Ugulino[1], Daniel Braun[2]
[1]University of Twente, [2]Marburg University
olegsilcenco@gmail.com, {m.r.machado,w.corbougulino}@utwente.nl, daniel.braun@uni-marburg.de

Abstract

Aspect-based sentiment analysis enhances sentiment detection by associating it with specific aspects, offering deeper insights than traditional sentiment analysis. This study introduces a manually annotated dataset of 10,814 multilingual customer reviews covering brick-and-mortar retail stores, labeled with eight aspect categories and their sentiment. Using this dataset, the performance of GPT-4 and LLaMA-3 in aspect based sentiment analysis is evaluated to establish a baseline for the newly introduced data. The results show both models achieving over 85% accuracy, while GPT-4 outperforms LLaMA-3 overall with regard to all relevant metrics.

1 Introduction

Sentiment analysis, i.e. the automatic identification of the sentiment expressed in, e.g., a text, is a widely used technique in research, business, politics, and many other domains (Wankhade et al., 2022). While traditional sentiment analysis methods focus mainly on detecting sentiment at the sentence or document level, aspect-based sentiment analysis is a more fine-grained approach through which particular aspects expressed in a text are identified with their corresponding sentiment (Zhang et al., 2022). For a movie review, e.g., in this way not just the overall sentiment expressed by the review is identified but also, for example, whether the reviewer liked or disliked the score or camera work. Such more fine-grained insights are particularly valuable to businesses as they provide better insights into the needs of customers.

While datasets for traditional sentiment analysis are widely available (see e.g. Tan et al. (2023); Kenyon-Dean et al. (2018); Wagh and Punde (2018), and Saif et al. (2013) for an overview of popular datasets), in part because they can be gathered in an automated fashion from services that use a combination of a score (e.g. in the form of stars) alongside with a textual review, the number of datasets available for aspect-based sentiment analysis is much more restricted (see e.g. Nazir et al. (2022) and Hua et al. (2024)). Additionally, most of the existing datasets either contain a small number of aspects per item or all aspects in one item have the same polarity (Jiang et al., 2019).

In this paper, we introduce a new, manually annotated, dataset for aspect-based sentiment analysis, that consists of 10,814 reviews for brick-and-mortar retail stores, scraped from Google Maps. The reviews cover different countries and languages and have been annotated with eight different aspect categories (see Table 3) resulting in a total of 16,994 labels. The dataset is available on GitHub[1]. A detailed datasheet (Gebru et al., 2021) for the corpus can be found in Appendix B.

In addition, we present a Large Language Model (LLM)-based baseline for the newly introduced dataset comparing the performance of Meta's LLaMa-3 and OpenAI's GPT-4. The results show that while both models perform well with an accuracy of more than 85%, GPT-4 consistently outperforms LLaMa-3 across aspects and metrics.

2 Related Work

Compared to "traditional" sentiment analysis, aspects-based sentiment analysis presents a more complex challenge, encompassing two distinct stages: identifying all aspects described, and subsequently determining the sentiment towards each aspect.

2.1 LLMs for Aspect-Based Sentiment Analysis

As for most NLP applications, recent literature about aspect-based sentiment analysis has mainly focused on the performance of LLMs. Recent

[1]https://github.com/Responsible-NLP/ABSA-Retail-Corpus

Table 1: Sample of scraped data

Country	City	Published At	Text	Stars
Belgium	Maasmechelen	18-03-2023	Najbolja i najkvalitetnija roba	5
France	Serris	29-12-2019	Great prices	5
Italy	Marcianise	21-11-2023	Ho scoperto questo negozio grazie...	5
Belgium	Mechelen	30-11-2017	Mooie propere zaak maar verkoper...	3

studies that compared the performance of LLMs against smaller language models (SLMs), like BERT, across a spectrum of sentiment analysis tasks, including conventional sentiment classification and aspect-based analysis, found that LLMs exhibit proficiency in simpler tasks, such as sentiment classification, but encounter difficulties in tasks requiring nuanced understanding or structured sentiment information (Zhang et al., 2024; Macháliková, 2023; Han and Moghaddam, 2023).

In few-shot learning scenarios, however, where annotation resources are limited, LLMs have shown superior performance (Zhang et al., 2024). According to Magdaleno et al. (2024), LLMs also outperform smaller models in such tasks as predicting ratings of businesses based on online reviews, and leveraging aspect-based sentiment analysis techniques. Moreover, LLMs have introduced innovative methodologies for context-aware analysis, as shown by Jeong and Lee (2024) in the context of hotel complaint reviews.

Comparative analyses between GPT-3.5, BERT, RoBERTa, and LLaMA report superior performance of GPT-3.5, specifically in predicting product review ratings post fine-tuning (Roumeliotis et al., 2024).

Krugmann and Hartmann (2024) report that using a zero-shot nature, LLMs can not only compete with but in some cases also surpass traditional transfer learning methods in terms of sentiment classification accuracy. Additionally, studies emphasize the competitive performance of GPT-3.5 model in discerning nuanced sentiments like irony within social media tweets, achieved through *prompt engineering* without explicit training (Carneros-Prado et al., 2023) Moreover, effective prompting engineering and *fine-tuning* are identified as crucial factors for achieving enhanced outputs and cost efficiency, further accentuating the potential of LLMs in customer satisfaction analysis and industry practices (Roumeliotis et al., 2024).

Comparative studies between LLMs and lexicon-based methods, show that LLMs clearly outperform such methods, while being particularly good in annotating sentiment analysis data and achieving over 94% accuracy in long-form sentiment reviews from Twitter social media users and Amazon customers, owing to its prowess in handling emojis, sarcasm, and contextual nuances (Belal et al., 2023). Notably, the literature suggests that GPT's integration into business customer sentiment analysis reveals its potential to significantly enhance understanding of customer sentiments, offering valuable insights for decision-making processes by comprehending both general sentiments and nuanced factors within customer texts (Sudirjo et al., 2023).

2.2 Datasets for Aspect-Based Sentiment Analysis

Table 2 shows a list of existing datasets for aspect-based sentiment analysis. While a number of datasets exists, the vast majority of them only covers the English language. Multilingual data sets, such as the one introduced in this paper, are particularly rare. With more than 10,000 annotated instances, the dataset introduced in this article is also one of the largest data sets for aspect-based sentiment analysis that is currently available.

3 Corpus

The data collection for the corpus relied on Google Maps reviews due to their accessibility and richness. While other publicly available datasets, such as those from product reviews or social media platforms, exist, many lack the level of granularity necessary for aspect-based sentiment analysis. These datasets typically emphasize general sentiment or aggregate ratings, which limits their suitability for examining specific aspects of customer feedback. Google Maps represent a robust platform where customers can leave reviews about a specific store or location. With its vast user base and widespread popularity, Google Maps is one of the most prominent platforms for user-generated reviews.

To efficiently scrape the reviews from Google

Table 2: Existing datasets for aspect-based sentiment analysis

Dataset	Domain	Lang.	Size	Sources
SemEval 2014	Service and Product Reviews	English	7,686	Pontiki et al. (2014)
SemEval 2015	Service and Product Reviews	English	5,596	Pontiki et al. (2015)
SemEval 2016	Service and Product Reviews	8	6,243	Pontiki et al. (2016)
Pars-ABSA	Service Reviews	Persian	5,602	Shangipour ataei et al. (2022)
Foursqaure	Service Reviews	English	585	Brun and Nikoulina (2018)
ACOS	Service and Product Reviews	English	6,362	Cai et al. (2021)
SentiHood	Neighbourhood Q&A	English	5,215	Saeidi et al. (2016)
MAMS	Service Reviews	English	13,854	Jiang et al. (2019)
Our dataset	Service Reviews	45	10,814	

Maps, the service Apify[2] was utilized. Privacy and personal data were primary concerns during the data collection process. Despite Google Maps reviews being publicly accessible, it was important to be cautious to ensure compliance with privacy standards. Apify's functionality enabled the selection of only the essential columns, omitting personal identifiers such as the name/nickname of the reviewer. Consequently, the collected data only included the review text, star rating, timestamp, and the location of the store (country and city). A small sample of such can be seen in Table 1.

3.1 Data Cleaning and Augmentation

A total of 24,361 reviews were collected in that way. Many reviews contained only a star rating without any textual review, these entries were excluded from the dataset, resulting in a final dataset of 10,814 reviews. Since the dataset contains reviews in a variety of languages, an additional column was added to encode the language of each review. The Google Translate API, accessed via the Python library `googletrans`, was utilized to detect the language of the reviews. The languages are encoded in ISO-639 format. The dataset also contains entries with unidentified languages and entries that contain only symbols or emojis. Additionally, the publication time, which was in textual format in the raw data, was converted to ISO-8601 format.

3.2 Data Annotation

The most important decision that had to be made before the data annotation is the definition of aspect categories. *Aspect categories* are higher-level concepts that pool different *aspect terms* to allow for more structured insights (Hua et al., 2024). The selection of the aspect categories was based on existing literature (particularly Kang et al. (2022); Ramaswami and Varghese (2003); Fakhira and Simanjuntak (2023)) and interviews with domain experts, to ensure that the chosen categories are relevant from both a scientific and a practitioner perspective. The eight aspect categories that have been derived from this process are shown in Table 3. They cover a broad range of customer experiences and provide valuable insights into different facets of the reviews and the businesses behind them. Each of these categories, if identified in a review, was assigned a sentiment label (negative, positive, or neutral).

The dataset was manually labeled by the authors. Manual labeling, though time-consuming, is critical for ensuring high-quality data annotations. Given the labor-intensive nature of this task, a custom labeling tool was developed to facilitate the process. Similar approaches have been employed in other studies. For instance, the SemEval-2014 task 4 involved creating annotation guidelines and tools for manual labeling to build benchmark datasets (Pontiki et al., 2014). Do et al. (2020) surveyed various tools and methods developed to assist in aspect-level sentiment annotation, while Li et al. (2012) presented a tool designed to create high-quality training data through manual annotation.

Figure 1 shows the tool that was developed to annotate the dataset. It allowed the authors to review the text, select the relevant aspect, and assign the corresponding sentiment. The tool features a user-friendly interface with functionalities such as language detection and translation to English, sentiment selection, and annotation saving.

10% of the data was annotated by two annotators to conduct an inter-annotator agreement study. Using Krippendorff's Alpha, the inter-annotator agree-

[2] https://apify.com/compass/google-maps-reviews-scraper

Table 3: Chosen aspect categories and their description

Aspect	Explanation
Product	Encompasses clothing collection, item quality, variety, display, and selection.
Service	Includes staff, assistance, crew, employee attitude, handling, and hospitality.
Brand	Pertains to overall brand perception.
Price	Relates to the cost of products and services, including promotions and discounts.
Store	Covers specific shop location, atmosphere, and environment.
Online	Concerns the online ordering experience.
Return	Includes the experience of returning an item for both physical or online procurement.
General	Overall experience of the customer without a specific aspect mention.

ment is $\alpha = 0.71$, a value that is comparable to other ABSA datasets and in general indicates a reliable annotation. An in-depth analysis revealed only two instances in which the annotators chose the same aspect but different polarities. In both cases it was not a direct contradiction, as in a positive and a negative label at the same time, but one label was neutral. An analysis on aspect-level shows that more specific aspects like price and service show higher agreement, while more general categories like store and "general" show lower agreement. A reoccurring pattern is that often, annotators agree on most aspects, but one annotator adds a single additional aspect (like general), thereby decreasing agreement.

3.3 Data Analysis

Out of the 10,814 reviews in the dataset, 4,838 (or 44.7%) contain more than one aspect. On average, each review contains 1.6 aspects. Figure 2 shows how often each aspect category occurs in the dataset. The most frequently occurring aspect category is service, which is mentioned in 6,065 reviews. The least mentioned category, online, only occurred in 51 reviews, which is not surprising, given that the reviews were specifically collected for physical store locations.

The majority of the aspects mentioned in the

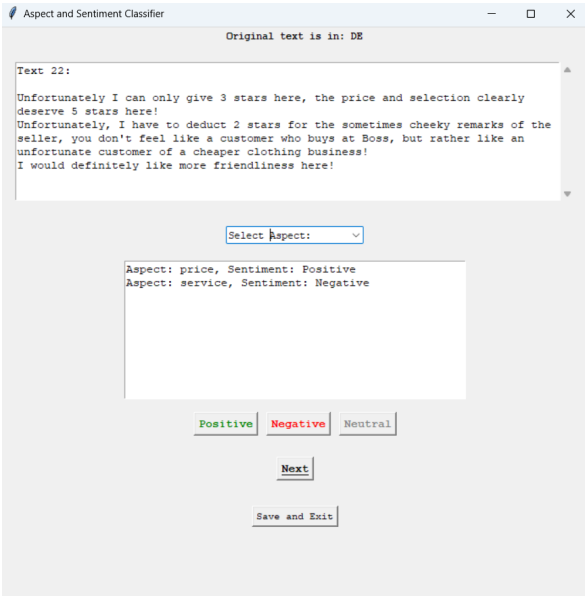

Figure 1: Labeling tool

dataset are positively connoted, as shown in Figure 3. Only for the aspect category return, the majority of the reviews expresses a negative sentiment. The most balanced aspect category is online, in which 55% of all mentioned aspects are positive, 8% neutral, and 37% negative. Neutral aspects are rarely mentioned across all categories. The highest share of neutral aspects can be found for the category general, with a little over 8%.

The reviews in the dataset have an average length of 121 characters, ranging from just one character (mostly emojis) to 3,735 characters. They cover stores from nine different European countries (Germany, France, Netherlands, Italy, Spain, Austria, Belgium, Portugal, and Switzerland; in descending frequency of occurrence) and are written in 45 different languages (see Figure 4 for a distribution of the languages).

4 Experimental Set-Up

To establish a baseline in aspect-based sentiment analysis for the newly introduced dataset, we conducted an experiment comparing the performance of the open weights model LLaMa-3 (Dubey et al., 2024) and the proprietary GPT-4 model (Achiam et al., 2023).

LLaMa-3 (`Meta-Llama-3-70B-Instruct`) was integrated through the HuggingFace library. To enhance performance and efficiency, quantization techniques were applied, reducing the model's weight precision via `BitsAndBytesConfig`, which optimized computational resources, particularly for

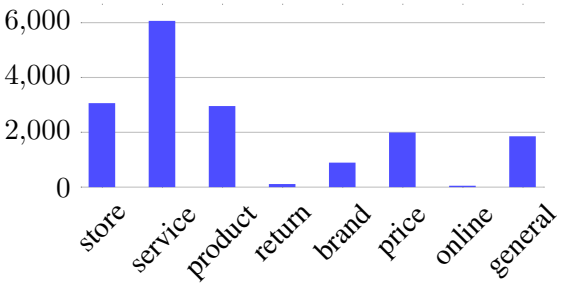

Figure 2: Occurrences of each aspect category

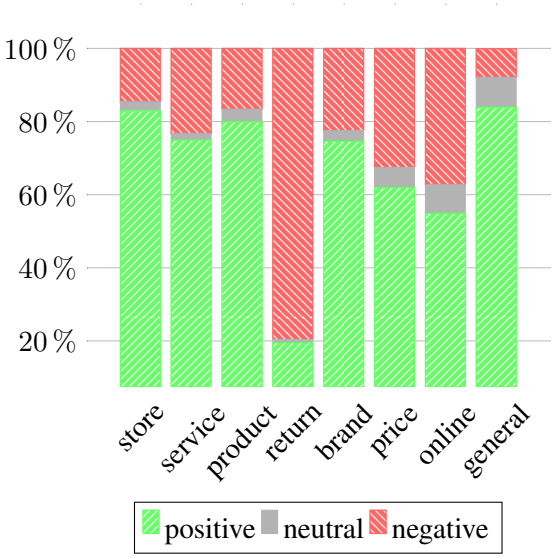

Figure 3: Share of positive, neutral, and negative aspects per category

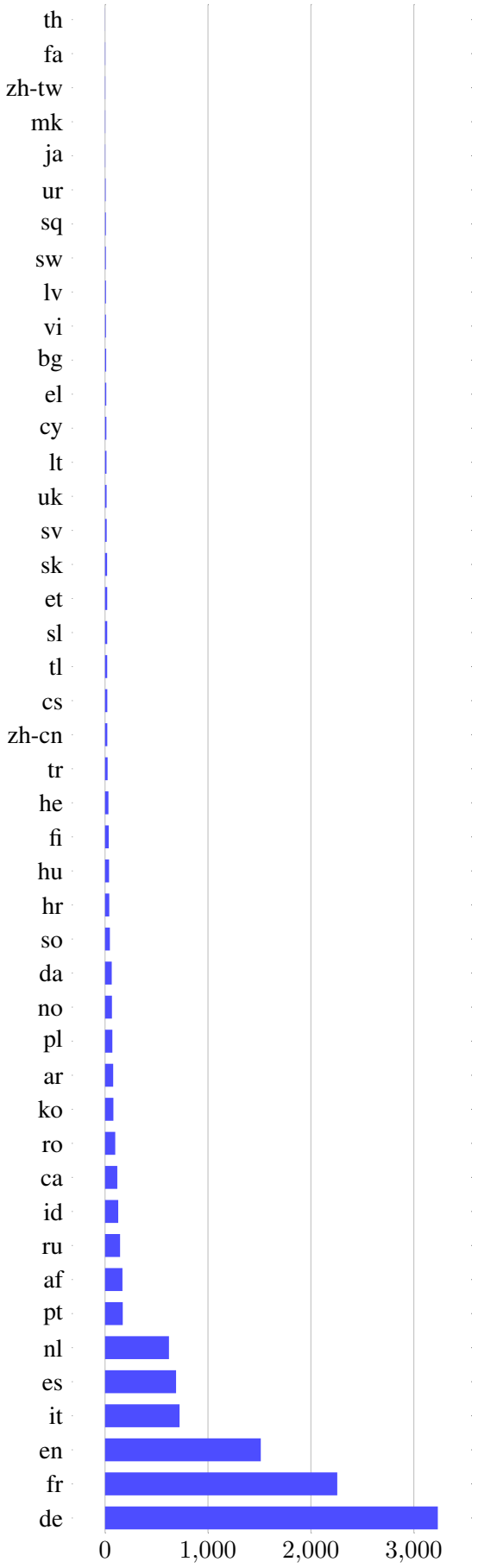

Figure 4: Number of reviews per language (ISO-639)

local deployment on GPU clusters.

GPT-4 was used through Azure's OpenAI Service. The implementation leveraged LangChain and its `PromptTemplate` component.

For both models, we used prompt engineering (Brown et al., 2020; Radford et al., 2019; Gao et al., 2021; Lu et al., 2021), and both system and user prompts, to facilitate the aspect-based sentiment analysis. *System prompts* are designed to define the role of the language model and establish operational guidelines to ensure consistency in responses. For both GPT-4 and LLaMA-3, the system prompt instructed the model to perform aspect-based sentiment analysis for a specific company. This approach is known as intent classification. The system prompt was implemented differently for each model, reflecting their respective frameworks. GPT-4 used LangChain's `SystemMessage` object to deliver the system instructions, while LLaMA-3 structured the system message as part of its chat template.

Despite syntactic differences, both implementations enforce a structured response format by defining system behavior upfront. LLaMA-3 specifies message roles such as *"system"* and *"user"* within its message list (see Listing 1), while GPT-4 utilizes LangChain's *langchain_core.messages* framework to differentiate between system and human messages (see Listing 2). These system prompts establish a consistent operational framework, ensuring the model generates precise and task-specific sentiment analysis responses.

User prompts provide the actual reviews and define the task parameters, guiding the model in identifying relevant aspects and classifying sentiments. The construction of these prompts is essential for ensuring accurate analysis, as they help the model distinguish between various aspects of a review and interpret sentiments effectively. To enhance performance, the prompts can incorporate structured instructions, few-shot examples, delimiters, and attention mechanisms.

One effective prompt engineering strategy is task decomposition, where a complex task is divided into smaller, more manageable steps. For instance:

> "First, identify the aspects in the provided review from the given list, and then find the customer sentiment (positive, neutral, or negative) for each of the aspects. ..."

Another key technique is few-shot prompting, where examples of correct outputs are included in the prompt to guide the model's response.

> " ... You can follow the examples below:
>
> Review: The product quality is great but the customer service is terrible. Aspect: product Sentiment: positive Aspect: service Sentiment: negative
>
> Review: I love the location of the store. The collection and selection look great, however, the prices are too high. Aspect: store Sentiment: positive Aspect: product Sentiment: positive Aspect: prices Sentiment: negative
>
> Review: All top, everything as I expected, recommend. Aspect: general Sentiment: positive ... "

Finally, attention mechanisms can be influenced by placing key instructions at the beginning and end of the prompt. For example:

> "First, identify the aspects in the provided review from the given list, and then find the customer sentiment (positive, neutral, or negative) for each aspect.
>
> Make sure to take into account the difference in language, cultural aspects, sarcasm, emojis, and other linguistic behaviors when interpreting and assessing the reviews. ...
>
> ... Remember to strictly focus only on the aspects from the list and reply only with the answer in the following JSON format:
>
> ["aspect1": "sentiment", ... "aspectN": "sentiment"]"

By placing important contextual elements at the beginning and specifying output format at the end, the model is guided to prioritize crucial information while maintaining structured responses. The final prompts used in the experiment can be found in Appendix A.

5 Results

Table 4 shows the evaluation of the aspect-based sentiment analysis experiment. Overall, GPT-4 outperformed LLaMA-3 in every single metric, with the widest gap in precision and the narrowest gap in recall. Given the imbalance between positive

Table 4: Precision, Recall, F1 Score, and Accuracy of the aspect-based sentiment analysis per model and aspect

	GPT-4				LLaMA-3			
Aspect	**Prec.**	**Recall**	**F1**	**Accur.**	**Prec.**	**Recall**	**F1**	**Accur.**
Store	**73.00%**	73.61%	**73.31%**	71.23%	59.02%	**80.13%**	67.97%	**75.53%**
Service	**93.88%**	**98.00%**	**95.89%**	**95.31%**	93.43%	96.44%	94.91%	94.05%
Product	**65.34%**	92.98%	**76.75%**	**88.39%**	58.12%	**93.70%**	71.74%	87.82%
Return	55.13%	**77.48%**	**64.42%**	**76.11%**	**62.89%**	54.46%	58.37%	53.98%
Brand	**57.92%**	70.65%	**63.66%**	65.66%	39.81%	**84.92%**	54.21%	**80.43%**
Price	**82.71%**	**85.86%**	**84.26%**	**79.23%**	79.04%	80.84%	79.93%	73.09%
Online	29.10%	**82.98%**	**43.09%**	**76.47%**	**30.09%**	72.34%	42.50%	66.67%
General	**49.75%**	**78.09%**	**60.78%**	**73.92%**	43.96%	65.13%	52.49%	62.13%
Micro avg.	**74.48%**	**87.58%**	**80.50%**	**83.80%**	66.84%	86.88%	75.55%	82.57%
Macro avg.	**63.35%**	**82.46%**	**70.27%**	**78.29%**	58.30%	78.49%	65.27%	74.21%

and negative reviews, the difference in accuracy of both models is only about 1.2 percentage points, despite the larger difference in precision. Only for the aspect categories “store” and “brand”, LLaMa-3 outperformed GPT-4 in respect to accuracy.

In high-frequency aspects such as “service” and “product”, both models showed strong performance, with GPT-4 recording accuracies of 95.31% and 88.39%, respectively, compared to LLaMA-3’s 93.05% and 87.82%. Notably, LLaMA-3 demonstrated improvements in the “product” category, narrowing the gap with GPT-4. These results indicate that both models reliably handle frequently mentioned aspects, although GPT-4 retains a slight edge in overall robustness.

When addressing lower-frequency or more nuanced aspects such as “return”, “brand”, and “online”, both models continued to face challenges, albeit with notable differences. GPT-4 demonstrated better performance in the “return” category, achieving an accuracy of 76.11% compared to LLaMA-3’s 53.98%. Similarly, in “online”, GPT-4 outperformed LLaMA-3, recording accuracies of 76.47% versus 66.67%. These findings underscore GPT-4’s greater capability to handle complex sentiment categories, though significant gaps remain. Both models struggled with the “online” aspect in terms of precision, with GPT-4 achieving 29.10% and LLaMA-3 slightly higher at 30.09%. These metrics highlight a broader limitation in capturing context-dependent nuances in less straightforward sentiment categories. In order for an aspect-based sentiment to be classified as correct, both the aspect and the sentiment expressed with it have to be extracted correctly. Notably, a large share of errors already occurs during the identification of aspects (see Table 5). The number of aspects that have been identified correctly but the sentiment was misclassified is relatively small.

6 Error Analysis

A deeper analysis of the errors made by both models revealed that LLaMA-3 exhibited a tendency to incorrectly identify aspects that are not present in the text based on mentioned keywords. For instance, in reviews such as “Great store for its amazing service and help from the assistants”, LLaMA-3 frequently identified “store” as an aspect, whereas the correct interpretation according to the aspect categories defined in Table 3 would be “service”. This over-sensitivity to mentions of keywords is also visible in Table 4, where the precision for the aspect store is particularly low for LLaMA-3. Yet, given the prevalence of the aspect across the dataset, this over-sensitivity might also partially explain why this is one of just two aspect categories in which LLaMA-3 achieved a higher recall and accuracy than GPT-4.

Both models encountered difficulties in consistently handling aspects from the categories “general” and “brand”. Reviews with broad or ambiguous sentiments about a brand in general, such as "Brand for Bosses," or "I love BOSS" for the clothing brand “BOSS” posed a significant challenge for both models. These reviews often led to inconsistent labeling, with models sometimes assigning both “brand” and “general” aspects or failing to distinguish between them altogether. Given the ambiguity of such statements, even human annotators would likely struggle to reach consensus, making this a particularly challenging area for automated analysis and labeling.

Table 5: Aspect identification accuracy

Aspect	GPT-4	LLaMA-3
Store	74.47%	**81.26%**
Service	**98.05%**	96.57%
Product	93.33%	**94.10%**
Return	**77.88%**	54.87%
Brand	72.73%	**85.71%**
Price	**86.95%**	82.69%
Online	**84.31%**	74.51%
General	**79.27%**	66.74%
Micro avg.	**88.12%**	87.57%
Macro avg.	**83.37%**	79.44%

LLaMA-3 showed a tendency to classify reviews under the apsect category "brand" more frequently than GPT-4, which contributed to its lower accuracy for the aspect category "general", scoring 66.74% comparing to 79.27% for GPT-4. Simultaneously, LLaMa-3 outperformed GPT-4 in the aspect category "brand", being one of only two aspect categories where it outperformed GPT-4 model. This behavior, while indicative of an attempt to capture a broader range of sentiment, often led to increased misclassification rates when the sentiment was intended to be more general. The higher labeling frequency for "brand" by LLaMA-3 also aligns with its overall lower precision and F1 scores in this aspect, reflecting a trade-off between recall and precision.

7 Conclusion

This paper introduced a new multilingual corpus for aspect-based sentiment analysis that is based on more than 10,000 reviews of brick-and-mortar stores and was manually labeled with eight aspect categories, namely product, service, brand, price, store, online, return, and general (see Table 3).

Additionaly, an experiment was conducted to establish a baseline for LLM-based aspect-based sentiment analysis on the newly introduced corpus by comparing the performance of GPT-4 and LLaMA-3. The results indicate that both models proficiently identify elements and attitudes in customer reviews, with GPT-4 continuously surpassing LLaMA-3 in precision, recall, and accuracy. Although both models achieved accuracy exceeding 85%, they performed insufficiently for the "store" and "brand" aspects, indicating areas for enhancement. From an application perspective, in addition to the performance, it is also worth considering the potential costs of different approaches and models. At the time of conducting the experiments, GPT-4 was queried through the Azure OpenAI API at a total cost of $240.60 for the complete dataset, or an average cost of $0.022 per review. With the price for one million input tokens being around $2.50 and the price for one million output tokens being around $10.00. LLaMA-3, on the other hand, cannot just be self hosted, but also used through cloud providers like groq who charge in the realm of $0.59 / $0.79 per one million input / output tokens, providing significantly cheaper options.

Ethics

By using public reviews and ensuring during both collection and annotation of the data that no identifiable information is contained in the reviews, we tried to ensure that our work has no direct adverse effects to anyone. Nevertheless, given the nature of the task, companies could use an approach like the one outlined in this work to try to automatically assess job performance of workers in physical store locations (e.g. by focusing on the aspect category service). However, we believe that in practice, the danger for such applications is relatively low given that most reviews (and none in our dataset) name specific employees or provide other information that could be used for the identification of individual employees like exact data and time of an interaction. While the environmental impact of LLMs is mostly discussed with regard to their training and fine-tuning, ever larger models also have an increasingly significant environmental impact during inference, which also holds true for this work.

Limitations

This study faced limitations inherent to the rapidly evolving field of LLM research. The introduction of newer models may render some aspects of this study's model selection less timely, although its fundamental methodologies remain relevant. Computational limitations also restricted fine-tuning and advanced quick engineering for LLaMA-3, potentially affecting its performance. Furthermore, the hand annotated dataset, albeit comprehensive, introduced subjectivity in aspect and sentiment classification, with ambiguous phrases and linguistic variances presenting hurdles to consistency. Lastly, the predefined aspect categories may not generalize across all use cases, making prompt design sensitive to initial definitions.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Mohammad Belal, James She, and Simon Wong. 2023. Leveraging chatgpt as text annotation tool for sentiment analysis. *arXiv preprint arXiv:2306.17177*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Caroline Brun and Vassilina Nikoulina. 2018. Aspect based sentiment analysis into the wild. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 116–122, Brussels, Belgium. Association for Computational Linguistics.

Hongjie Cai, Rui Xia, and Jianfei Yu. 2021. Aspect-category-opinion-sentiment quadruple extraction with implicit aspects and opinions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 340–350, Online. Association for Computational Linguistics.

David Carneros-Prado, Laura Villa, Esperanza Johnson, Cosmin C Dobrescu, Alfonso Barragán, and Beatriz García-Martínez. 2023. Comparative study of large language models as emotion and sentiment analysis systems: A case-specific analysis of gpt vs. ibm watson. In *International Conference on Ubiquitous Computing and Ambient Intelligence*, pages 229–239. Springer.

Bao Lieu Do, Lam Pham Huy, and Xuan-Linh Tran. 2020. Automated tools for aspect-level sentiment analysis: A survey. *Information Processing & Management*, 57(6):102311.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Nisrina Nur Fakhira and Megawati Simanjuntak. 2023. Content analysis of consumer reviews and comments on e-commerce. *Jurnal Doktor Manajemen*, 6:2.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III au2, and Kate Crawford. 2021. Datasheets for datasets. *Preprint*, arXiv:1803.09010.

Yi Han and Mohsen Moghaddam. 2023. A design knowledge guided position encoding methodology for implicit need identification from user reviews. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, volume 87295, page V002T02A092. American Society of Mechanical Engineers.

Yan Cathy Hua, Paul Denny, Jörg Wicker, and Katerina Taskova. 2024. A systematic review of aspect-based sentiment analysis: domains, methods, and trends. *Artificial Intelligence Review*, 57(11):296.

Nayoung Jeong and Jihwan Lee. 2024. An aspect-based review analysis using chatgpt for the exploration of hotel service failures. *Sustainability*, 16(4):1640.

Qingnan Jiang, Lei Chen, Ruifeng Xu, Xiang Ao, and Min Yang. 2019. A challenge dataset and effective models for aspect-based sentiment analysis. 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 6280–6285, Hong Kong, China. Association for Computational Linguistics.

Min Kang, Bing Sun, Tian Liang, and Hong-Ying Mao. 2022. A study on the influence of online reviews of new products on consumers' purchase decisions: An empirical study on jd. com. *Frontiers in Psychology*, 13:983060.

Kian Kenyon-Dean, Eisha Ahmed, Scott Fujimoto, Jeremy Georges-Filteau, Christopher Glasz, Barleen Kaur, Auguste Lalande, Shruti Bhanderi, Robert Belfer, Nirmal Kanagasabai, Roman Sarrazingendron, Rohit Verma, and Derek Ruths. 2018. Sentiment analysis: It's complicated! In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1886–1895, New Orleans, Louisiana. Association for Computational Linguistics.

Jan Ole Krugmann and Jochen Hartmann. 2024. Sentiment analysis in the age of generative ai. *Customer Needs and Solutions*, 11(1):3.

Fang Li, Yulan He, Weizhi Meng, and Kun Wu. 2012. A generic approach for sentiment analysis and opinion mining. *International Journal of Computer Applications*, 49(3):21–28.

Xinxi Lu, Antoine Bosselut, Junyi Jessy Hao, and Yejin Choi. 2021. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4074–4088.

Kristina Macháliková. 2023. Utilizing chatgpt for sentiment analysis.

Diego Magdaleno, Martin Montes, Blanca Estrada, and Alberto Ochoa-Zezzatti. 2024. A gpt-based approach for sentiment analysis and bakery rating prediction. In *Advances in Computational Intelligence. MICAI 2023 International Workshops*, pages 61–76, Cham. Springer Nature Switzerland.

Ambreen Nazir, Yuan Rao, Lianwei Wu, and Ling Sun. 2022. Issues and challenges of aspect-based sentiment analysis: A comprehensive survey. *IEEE Transactions on Affective Computing*, 13(2):845–863.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. SemEval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495, Denver, Colorado. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Seshan Ramaswami and Susheela Abraham Varghese. 2003. Reading the voice of the customer: A content analysis of consumer reviews.

Konstantinos I Roumeliotis, Nikolaos D Tselikas, and Dimitrios K Nasiopoulos. 2024. Llms in e-commerce: a comparative analysis of gpt and llama models in product review evaluation. *Natural Language Processing Journal*, 6:100056.

Marzieh Saeidi, Guillaume Bouchard, Maria Liakata, and Sebastian Riedel. 2016. SentiHood: Targeted aspect based sentiment analysis dataset for urban neighbourhoods. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1546–1556, Osaka, Japan. The COLING 2016 Organizing Committee.

Hassan Saif, Miriam Fernández, Yulan He, and Harith Alani. 2013. Evaluation datasets for twitter sentiment analysis: a survey and a new dataset, the sts-gold. In *1st Interantional Workshop on Emotion and Sentiment in Social and Expressive Media: Approaches and Perspectives from AI (ESSEM 2013)*.

Taha Shangipour ataei, Kamyar Darvishi, Soroush Javdan, Behrouz Minaei-Bidgoli, and Sauleh Eetemadi. 2022. Pars-ABSA: a manually annotated aspect-based sentiment analysis benchmark on Farsi product reviews. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7056–7060, Marseille, France. European Language Resources Association.

Frans Sudirjo, Karno Diantoro, Jassim Ahmad Al-Gasawneh, Hizbul Khootimah Azzaakiyyah, and Abu Muna Almaududi Ausat. 2023. Application of chatgpt in improving customer sentiment analysis for businesses. *Jurnal Teknologi Dan Sistem Informasi Bisnis*, 5(3):283–288.

Kian Long Tan, Chin Poo Lee, and Kian Ming Lim. 2023. A survey of sentiment analysis: Approaches, datasets, and future research. *Applied Sciences*, 13(7).

Rasika Wagh and Payal Punde. 2018. Survey on sentiment analysis using twitter dataset. In *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pages 208–211.

Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. 2022. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7):5731–5780.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Pan, and Lidong Bing. 2024. Sentiment analysis in the era of large language models: A reality check. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3881–3906, Mexico City, Mexico. Association for Computational Linguistics.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2022. A survey on aspect-based sentiment analysis: Tasks, methods, and challenges. *IEEE Trans. on Knowl. and Data Eng.*, 35(11):11019–11038.

A Prompts

Listing 1: LLaMA-3 system prompt

```
messages = [ {"role": "system", "content
    ": """You are a helpful assistant
    that performs aspect based sentiment
     analysis for Hugo Boss! Do not
    communicate back, just provide the
    answer in the requested format"""},
    {"role": "user", "content": text}, ]
prompt = output.tokenizer.
    apply_chat_template( messages,
    tokenize=False,
    add_generation_prompt=True )
```

Listing 2: GPT-4 system prompt

```
system_prompt = """You are a helpful
    assistant that performs aspect based
     sentiment analysis for Hugo Boss!
    Do not communicate back, just
    provide the answer in the requested
    format."""

prompt_value = StringPromptValue(text=
    chat_prompt_with_values)

output = llm.invoke([ SystemMessage(
    content=system_prompt), HumanMessage
    (content=prompt_value.text), ])
```

Listing 3: LLaMA-3 user prompt

```
text = f"""First, identify the aspects
    in the provided review from the
    given list, and then find the
    customer sentiment (positive,
    neutral, or negative) for each of
    the aspect.
Make sure to take into account the
    difference in the language, cultural
     aspects, sarcasm, emojis, and other
     linguistic behaviours when
    interpreting and assessing the
    reviews.

Aspects: [Product (collection, item,
    quality, variety, display, selection
    ), Service (staff, assistance, crew,
     employee, attitude, handling,
    hospitality), Brand, Price, Store (
    shop, location, atmosphere), Online
    (order), Purchase, Return, General (
    overall shopping experience)]

You can follow the examples below:

Review: The product quality is great but
     the customer service is terrible.
Aspect: product
Sentiment: positive
Aspect: service
Sentiment: negative

Review: I love the location of the store
    . The collection and selection looks
     great, however the prices are too
    high.
Aspect: store
Sentiment: positive
Aspect: product
Sentiment: positive
Aspect: prices
Sentiment: negative

Review: All top, everything as I
    expected, recommend.
Aspect: general
Sentiment: positive

Now proceed with the following review:
    ```{review}```

Remember to strictly focus only on the
    aspects from the list and reply only
     with the answer in the following
    JSON format:
[{{"aspect1": "sentiment"}},
...
{{"aspectN": "sentiment"}}]
"""
```

Listing 4: GPT-4 user prompt

```
First, identify the aspects in the
    provided review from the given list,
     and then find the customer
    sentiment (positive, neutral ot
    negative) for each of the aspect.

Aspects: [Product (collection, item,
    quality, variety, display), Service
    (staff, assistance, crew, employee,
    attitude, handling, hospitality),
    Brand, Price, Store (shop, location,
     atmosphere), Online (order), Return
    , General (overall shopping
    experience)]

You can follow the examples below:
review: "The␣product␣quality␣is␣great␣
    but␣the␣customer␣service␣is␣terrible
    ."
analysis: "[{{"product":␣"positive",␣"
    service":␣"negative"}}]"

review: "I␣love␣the␣location␣of␣the␣
    store.␣The␣collection␣looks␣great,␣
    however␣the␣prices␣are␣too␣high.",
analysis: "[{{"store":␣"positive",␣"
    product":␣"positive",␣"price":␣"
    negative"}}]"

review: "All␣top,␣everything␣as␣I␣
    expected,␣recommend.",
analysis: "[{{"general":␣"positive"}}]"

Make sure to take in account the
    difference in the language, cultural
     aspects, sarcasm, emojis and other
    linguistic behaviours when
    interpreting and assessing the
    reviews. Remember to strictly reply
    only with the answer in the
    following JSON format:
[{{"aspect1": "sentiment"}},
...
{{"aspectN": "sentiment"}}]
```
```

B Datasheet

B.1 Motivation for Dataset Creation

Why was the dataset created? (e.g., were there specific tasks in mind, or a specific gap that needed to be filled?)

The dataset was created to enable aspect-based sentiment analysis on customer reviews using Large Language Models (LLMs).

What (other) tasks could the dataset be used for? Are there obvious tasks for which it should not be used?

The dataset could also be used for traditional sentiment analysis given it also contains star ratings for each review.

Has the dataset been used for any tasks already? If so, where are the results so others can compare (e.g., links to published papers)?

This paper is the first to use the dataset.

Who funded the creation of the dataset? If there is an associated grant, provide the grant number.

The data collection was supported by the consulting firm Metyis (`https://metyis.com/`).

B.2 Dataset Composition

What are the instances? (that is, examples; e.g., documents, images, people, countries) Are there multiple types of instances? (e.g., movies, users, ratings; people, interactions between them; nodes, edges)

Each instance consists of a customer review for a brick-and-mortar store, scarped from Google maps.

Are relationships between instances made explicit in the data (e.g., social network links, user/-movie ratings, etc.)?

No.

How many instances of each type are there?

The dataset consists of 10,814 reviews.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images)? Features/attributes? Is there a label/target associated with instances? If the instances are related to people, are subpopulations identified (e.g., by age, gender, etc.) and what is their distribution?

In addition to the country and city of the store that is reviewed, the date the review was published at, its text, and the star rating are part of each instance. The language of the review is automatically annotated, while aspects and their sentiments have been manually annotated.

Is everything included or does the data rely on external resources? (e.g., websites, tweets, datasets) If external resources, a) are there guarantees that they will exist, and remain constant, over time; b) is there an official archival version. Are there licenses, fees or rights associated with any of the data?

Everything is included in the dataset.

Are there recommended data splits or evaluation measures? (e.g., training, development, testing; accuracy/AUC)

Since the dataset is designed for zero-shot classification, there is no recommended split. Given the imbalanced distribution of positive and negative sentiments, we recommend and evaluation measure that takes this into account, like F1-score.

What experiments were initially run on this dataset? Have a summary of those results and, if available, provide the link to a paper with more information here.

The dataset was initially used for the evaluation of the aspect-based sentiment analysis capabilities of LLMs.

B.3 Data Collection Process

How was the data collected? (e.g., hardware apparatus/sensor, manual human curation, software program, software interface/API; how were these constructs/measures/methods validated?)

The data was collected using Apify (`https://apify.com/compass/google-maps-reviews-scraper`).

Who was involved in the data collection process? (e.g., students, crowdworkers) How were they compensated? (e.g., how much were crowdworkers paid?)

The data was collected by fully-qualified lawyers during their usual work-time. All participants worked for organizations that pay according to the collective labor agreement for public service workers in German states.

Over what time-frame was the data collected? Does the collection time-frame match the creation time-frame?

The data was collected in 2024. The reviews were written between 2012 and 2024.

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part of speech tags; model-based guesses for age or language)? If

the latter two, were they validated/verified and if so how?

The reviews themselves and their metadata was directly observable, the language of the reviews was automatically derived using the Google Translate API, the aspects and their sentiments were manually annotated by the authors.

Does the dataset contain all possible instances? Or is it, for instance, a sample (not necessarily random) from a larger set of instances?

No, the dataset does not claim completeness in any sense.

If the dataset is a sample, then what is the population? What was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)? Is the sample representative of the larger set (e.g., geographic coverage)? If not, why not (e.g., to cover a more diverse range of instances)? How does this affect possible uses?

The dataset spans multiple European countries and a time-frame of over a decade.

Is there information missing from the dataset and why? (this does not include intentionally dropped instances; it might include, e.g., redacted text, withheld documents) Is this data missing because it was unavailable?

Reviews that only consists of a star rating but do not provide any text have been excluded.

B.4 Dataset Distribution

How is the dataset distributed? (e.g., website, API, etc.; does the data have a DOI; is it archived redundantly?)

It is archived on GitHub (`https://github.com/Responsible-NLP/ABSA-Retail-Corpus`).

When will the dataset be released/first distributed? (Is there a canonical paper/reference for this dataset?)

Publication of the paper.

What license (if any) is it distributed under? Are there any copyrights on the data?

The annotations are licensed under CC-BY-SA 4.0.

Are there any fees or access/export restrictions?

No.

B.5 Dataset Maintenance

Who is supporting/hosting/maintaining the dataset? How does one contact the owner/curator/manager of the dataset (e.g. email address, or other contact info)?

See the GitHub repository.

Will the dataset be updated? How often and by whom? How will updates/revisions be documented and communicated (e.g., mailing list, GitHub)? Is there an erratum?

There are no plans to update the dataset unless important mistakes become clear.

If the dataset becomes obsolete how will this be communicated?

On the GitHub page.

Is there a repository to link to any/all papers/systems that use this dataset?

Yes.

If others want to extend/augment/build on this dataset, is there a mechanism for them to do so? If so, is there a process for tracking/assessing the quality of those contributions. What is the process for communicating/distributing these contributions to users?

We would suggest to create a fork on GitHub.

B.6 Legal & Ethical Considerations

If the dataset relates to people (e.g., their attributes) or was generated by people, were they informed about the data collection? (e.g., datasets that collect writing, photos, interactions, transactions, etc.)

There is no information about individuals in the data or was recorded during the annotation of the data.

If it relates to other ethically protected subjects, have appropriate obligations been met? (e.g., medical data might include information collected from animals)

N.a.

If it relates to people, were there any ethical review applications/reviews/approvals? (e.g. Institutional Review Board applications)

N.a.

If it relates to people, were they told what the dataset would be used for and did they consent? What community norms exist for data collected from human communications? If consent was obtained, how? Were the people provided with any mechanism to revoke their consent in the future or for certain uses?

N.a.

If it relates to people, could this dataset expose people to harm or legal action? (e.g., financial social or otherwise) What was done to mitigate or reduce the potential for harm?

N.a.

If it relates to people, does it unfairly advantage or disadvantage a particular social group? In what ways? How was this mitigated?

N.a.

If it relates to people, were they provided with privacy guarantees? If so, what guarantees and how are these ensured?

N.a.

Does the dataset comply with the EU General Data Protection Regulation (GDPR)? Does it comply with any other standards, such as the US Equal Employment Opportunity Act?

Yes, since only publicly available information was collected, the dataset complies with the GDPR and similar regulations.

Does the dataset contain information that might be considered sensitive or confidential? (e.g., personally identifying information)

No.

Does the dataset contain information that might be considered inappropriate or offensive?

No.