A GPT among Annotators: LLM-based Entity-Level Sentiment Annotation

Egil Rønningstad and **Erik Velldal** and **Lilja Øvrelid**
University of Oslo, Language Technology Group
{egilron,erikve,liljao}@ifi.uio.no

Abstract

We investigate annotator variation for the novel task of Entity-Level Sentiment Analysis (ELSA) which annotates the aggregated sentiment directed towards volitional entities in a text. More specifically, we analyze the annotations of a newly constructed Norwegian ELSA dataset and release additional data with each annotator's labels for the 247 entities in the dataset's test split. We also perform a number of experiments prompting ChatGPT for these sentiment labels regarding each entity in the text and compare the generated annotations with the human labels. Cohen's Kappa for agreement between the best LLM-generated labels and curated gold was 0.425, which indicates that these labels would not have high quality. Our analyses further investigate the errors that ChatGPT outputs, and compare them with the variations that we find among the 5 trained annotators that all annotated the same test data.

1 Introduction

Sentiment analysis constitutes an annotation task that is highly subjective in nature, typically with moderate inter-annotator agreement levels (Bobicev and Sokolova, 2017; Provoost et al., 2019; Kim and Klinger, 2018; Barnes et al., 2021). Recently, Gilardi et al. (2023) and Alizadeh et al. (2023) compare the label quality of crowd workers to the labels generated by a large language model (LLM). They show how annotations by a LLM can surpass the quality obtained from crowd workers, for certain annotation tasks for the English language. Šmíd and Přibáň (2023) employ multilingual generative language models (mT5 and mBART) for other sentiment analysis tasks through prompts and fine-tuning for Czech sentiment analysis. They find that prompting these types of models can be a potentially promising avenue for few-shot or zero-shot scenarios.

An important direction in recent work on sentiment analysis focuses on analysis of longer texts (Dufraisse et al., 2023; Rønningstad et al., 2022). Our task of Entity-Level Sentiment Analysis (ELSA) follows in this direction. It was introduced and motivated by Rønningstad et al. (2022) and can be defined as providing one sentiment score for each volitional entity in a text, the reader's total impression from reading the entire text. The overall, entity-wise sentiment is central to this work, since the annotated texts are quite long, containing several different entities, whereby each entity may be referenced in several sentences in separate parts of the text. A volitional entity, or just "entity" in our context, is a person or organization mentioned by its proper name in the text. We recently released a manually annotated dataset for the ELSA task where the overall sentiment conveyed in the text towards each volitional entity is annotated.[1] Figure 1 shows an example text with ELSA annotations.

In this paper we present a number of experiments attempting to generate ELSA sentiment annotations by prompting ChatGPT, and we perform an in-depth comparison between the LLM-generated labels, and the labels provided by five human annotators that all labeled the test-split of our dataset in parallel. With this paper we also make available the annotators' labels for each entity in the test set. We believe that the ELSA annotation task offers a challenging testbed for LLM-based annotation due to the following characteristics:

Longer texts The texts are professional published reviews with a mean sentence count of 27.5.

Norwegian language The texts are in the Norwegian language, a small language which amounts to a minuscule portion of the GPT pre-training data;

[1] https://github.com/ltgoslo/ELSA

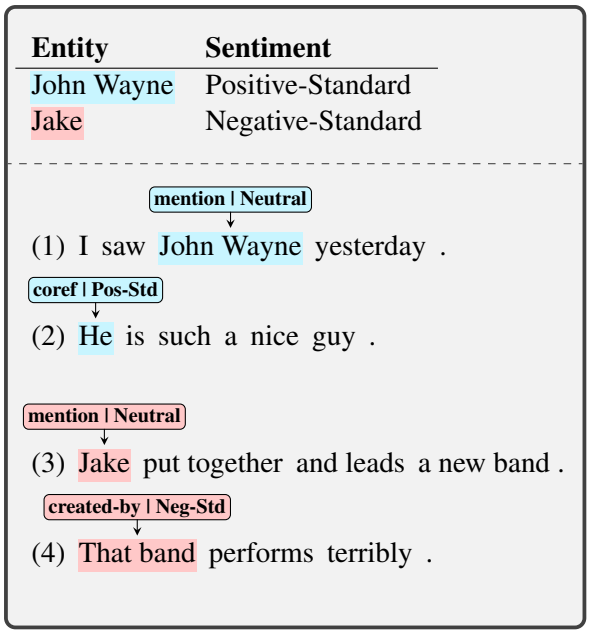

Figure 1: Toy example of one text containing two entities and their overall sentiment classification (top), together with their references in the text, with sentiment classifications. The sentiments here are not expressed directly towards the entity mention, but towards a coreference in sentence 2 and towards something "created by" the entity in question, in sentence 4. The reader considers sentence 4 to convey a negative sentiment with respect to the entity "Jake", since he appears to be so central to it.

New task definition The specific sentiment analysis task (ELSA) is to the best of our knowledge not annotated in other publicly available datasets that could have been seen during pretraining;

Long distance relations The task requires filtering sentiment expressions with regards to the entity in question, to connect all relevant expressions of the document and aggregate this into a sentiment label from the provided set.

2 ELSA annotation

The newly created ELSA dataset contains the texts of the NoReC$_{fine}$ dataset (Øvrelid et al., 2020) – a subset of the Norwegian Review Corpus (NoReC; (Velldal et al., 2018)) of professional reviews from a variety of domains, e.g. screen, music, restaurants and literature. The ELSA annotations adds sentiment information at the sentence- and document-level for each volitional entity in a given text; entities of the types Person (PER) and Organization (ORG). The texts contain on average mentions of 6 different entities, and each entity is assigned an individual sentiment label based on the reader's overall impression from the text. Sentiment polarity is classified with two intensity levels, providing a set of five possible labels: "Positive-Standard", "Positive-Slight", "Neutral", "Negative-Slight", and "Negative-Standard".

The ELSA dataset is annotated by 5 annotators, all native Norwegian undergraduate NLP-students. After introductory training and fine-tuning of the guidelines, the annotators (single-)annotated separate parts of the main body of the dataset, i.e. the training and development splits. The test set, consisting of 44 documents with a total of 1252 sentences and annotated sentiment toward 247 unique entities, was subsequently annotated by all five annotators. The entire dataset was manually curated by the project leader. Curation consisted mainly of error-correction and assigning the majority vote, which was accepted for 90% of the entities. An English translation of the annotator guidelines is found in the supplementary materials.

2.1 Individual variations

Since all five annotators labeled the sentiment regarding all 247 entities in the test set, we can study the variation between well-informed human readers regarding perceived sentiment presented in a longer text. The bottom rows of Table 2 shows the variations between annotators and the curated version. We consider the views and findings of Pavlick and Kwiatkowski (2019) to be relevant when studying the inter-annotator agreement in our dataset. They find that for Natural Language Inference annotations, it is hard to support the view that disagreement between annotators should be dismissed as annotation "noise". For our dataset, we do consider the manually expert-curated gold to be the best available single representation of the sentiment expressed in the texts regarding each entity. But we also consider each annotator's labels to represent a valid reading of the text and of that annotator's classification of the sentiment perceived. We find that the average κ score is .68 which can be characterised as moderate/good, however, also observe that there is considerable variation between the annotators, and in particular one annotator (ann_1) represents an outlier with diverging annotations.

3 LLM-based annotation

We employ OpenAI's ChatGPT (OpenAI, 2023) in order to explore whether a widely used LLM can

provide sentiment labels for volitional entities in longer Norwegian texts, similar to that of native human annotators. We here present the alternatives explored in our experiments. The results from the various experiments can be found in Table 1.

The model prompted was GPT-4, `gpt-4-1106-preview`. For all our experiments, we instructed the model to output both a justification regarding the sentiment of each entity, and to provide the concluded labels separately in structured format (JSON). The prompts used in the experiment yielding best results can be found in the appendix. The instructions, prompts and replies for one text during the earlier "gpt06" experiment, can be found in the supplementary material.

Norwegian or English prompting We experimented with both Norwegian and English prompts. We always specified in the instructions that the texts would be in Norwegian, and we mostly got Norwegian text back.

Requesting one or many analyses per prompt As mentioned above, each document contains on average six entities to label. We experimented with either listing all entities to label in one prompt per document, but also in one case to submit one prompt per entity. The chosen variant is shown in the "Entities per prompt" column in Table 1.

Adding knowledge For our two last experiments, gpt0801 and gpt0802, we experimented with the addition of more information to the system instructions. For gpt0801, a condensed version of the Norwegian annotation guidelines were uploaded and referenced in the system instructions. For gpt0802, we additionally added a text file containing three texts, with their entities and each entity's sentiment label annotated, thus corresponding to a few-shot scenario.

4 Findings

Table 1 presents our experiments and the evaluation results for the labels generated by ChatGPT. We here discuss the impact from alternating the options listed in Table 1, and our analysis of the labels generated by the best performing setup.

4.1 Best model

The annotations resulting from the experiment "gpt0802", i.e. the few-shot scenario described above, shows the best accuracy and weighted F_1, measured against the curated gold standard, and we choose these results for further analyses. We find that only 62.8% of the best GPT-generated labels are correct. The Cohen's Kappa (κ) agreement is 42.5%. When we compare with the mean values for the five annotators, we find that the GPT-generated labels perform noticeably poorer than the annotators' average. However, when we compare the GPT-generated labels with each of the five annotators' scores, we find that, by some metrics, one annotator deviates more from the curated gold than the GPT output does.

4.2 Adding knowledge helps

From our results in Table 1 we see that the two final experiments, gpt0801 and gpt0802, resulted in better annotations than the previous experiments. One example where the outputs in gpt0802 were accurate, while earlier experiments produced an incorrect label is found in a movie review where the entity "Hitler" is mentioned. The reference is used to place the the events of the movie in space and time, during the last days of Hitler. The text does not attribute the horrors of war directly towards this entity. Previous GPT-experiments yielded the label "Negative-Standard", while the label for "Hitler" in gpt0802 is "Neutral". The curated gold value for the entity is "Neutral". We speculate that the model tends to employ too much previous knowledge about the entity from its pretraining, while with the more thorough instructions given in the annotators' guidelines and in the examples, the model aligned to this information and yielded a label derived more from the text in question only.

4.3 Evaluation against curated

The lower part of Table 2 shows the gap between GPT-generated labels and annotators' average in terms of accuracy, κ and Mean Squared Error (MSE). MSE is calculated using the numerical mappings for the sentiment labels provided in Table 3, and indicates the distance with which a label deviates from the curated gold. We find that the generated labels are generally further away from the true labels, once again with the exception of the outlier annotator.

4.4 Majority, Within range or Outside

An interesting question in the current context relates to whether the LLM-based annotation errors are qualitatively different than the observed annotator variation. In order to assess this we computed

Configuration	Prompt language	Entities per prompt	Accuracy	W. F1	System files
gpt06	Norwegian	document	0.53	0.6	
gpt07	English	individual	0.49	0.54	
gpt0701	English	document	0.37	0.41	
gpt0801	Norwegian	document	0.58	0.61	guidelines
gpt0802	Norwegian	document	0.63	0.65	guidelines + 3 examples

Table 1: Our experimental setups with accuracy and weighted F_1 measured against curated gold data. Section 3 explains the various design options. Under "Entities per prompt", "document" indicates that all entities in one document were submitted in one prompt. "individual" means we submitted one prompt per entity in the document. The "Configuration" columns apply the working title for each experiment as identification. The five first experiments were considered to be introductory and are not reported on.

	ann_avg	gpt	ann_1	ann_2	ann_3	ann_4	ann_5	curated
Majority	0.816	0.623	0.623	0.895	0.806	0.842	0.915	0.907
Within range	0.089	0.219	0.117	0.061	0.121	0.089	0.057	0.093
Outside	0.095	0.158	0.259	0.045	0.073	0.069	0.028	0.000
Accuracy	0.804	0.628	0.684	0.862	0.818	0.789	0.866	1.000
κ	0.683	0.425	0.543	0.764	0.712	0.625	0.771	1.000
MSE	0.144	0.220	0.237	0.103	0.096	0.165	0.116	0.000

Table 2: For the upper part of the table, we have counted for each label assigned by the annotators, whether it agrees with the majority of annotators, within the label span created by the other annotators, or outside this span. Further discussion of these values is found in Section 4.4. The lower part of the table shows Accuracy, Cohen's Kappa, and Mean Squared Error for each annotator, measured against curated.

Label	Numerical
Negative-Standard	-1.0
Negative-Slight	-0.5
Neutral	0.0
Positive-Slight	0.5
Positive-Standard	1.0

Table 3: Numerical mapping for the sentiment labels, in order to calculate mean square errors.

for each of the annotators, the one or two labels assigned by a majority of the four other annotators. If the label assigned by the annotator in question equaled such a majority vote, the label was counted as "Majority". If not, we examined the maximum and minimum value for the labels assigned by the other annotators, using the conversion table in Table 3. If the label was within the range defined by the labels of the other annotators, we counted this label as "Within range". The labels that were outside the labels range of other annotators, were counted as "Outside". The results are shown in the upper part of Table 2. Note that each annotator was evaluated by the 4 other annotators, while the GPT-generated labels were evaluated against the labels from all 5 annotators. We see in Table 2 for the row "Outside", that 15.8% of the labels generated by GPT was deviating more from the majority vote than any annotator. This is more than 50% more than the annotator average (9.5%), but not more than for the outlier Annotator 1 (25.9%).

4.5 Post-processing GPT outputs

We find in general that the generated texts from our interaction with GPT are not always consistent with the instructions, neither in repeating the entity name, in the labels assigned, nor in the formatting of the requested JSON output. After implementing a post-processing heuristics for extracting the entity name and sentiment label in accordance with the standard GPT reply, we aligned the GPT output with our annotated data by inspecting each format error and creating a rule-based conversion script. Each experiment with its variations in system instructions and prompt wordings, yielded different format variations in the output. For the given dataset with 247 entities, this post-processing was

manageable. For a larger dataset, tested for more variations in the setup, this will be a non-negligible part of the task.

5 Suggestions for future work

From the findings reported in this paper, we are personally not encouraged to undertake new annotation projects using LLM-based annotation only. But we see a potential for augmenting datasets using a similar approach. When it comes to modelling, both the ELSA dataset and other SA datasets may benefit from being extended by LLM annotations. Including more open source instruction-tuned LLMs would then be essential. Modelling with LLMs is also highly relevant, and the experiments reported here can be used as the starting point for further experiments in that direction.

6 Conclusions

We have presented and analyzed a dataset of annotators' agreement for the task of Entity-Level Sentiment Analysis in Norwegian, and studied how well GPT-generated labels compare with labels produced by five different human annotators for this particular task. We consider the ELSA task to be interesting for such comparisons since the entire text needs to be analyzed for each entity, and the texts are lengthy, varying around a mean of 27.5 sentences per text. We find that the GPT-generated labels have lower accuracy and Cohen's Kappa than any annotator. But when we try to quantify the magnitude of the deviations, *e. g.* through MSE, we find that one "outlier" annotator has stronger deviations from gold that the GPT-generated labels. This represents new insights for us, since it has been assumed that the errors made by an llm could be more dramatic than those made by humans.

Similar approaches may therefore be well worth exploring for other sentiment analysis tasks in other languages similarly related to English. Particularly in parallel with human annotators. The only method we are aware of which serves to explore how a LLM could help with a certain new task, is by annotating data manually, since the task is to identify the sentiment as perceived by humans.

7 Limitations

7.1 Norwegian Language

We have analyzed a Norwegian dataset. We find in general that ChatGPT does a decent job in translating between Norwegian and English. We believe that the ChatGPT performance we found, might be matched or surpassed by other languages related to English, *e. g.* Germanic or Italic languages, with a similar or stronger web presence than Norwegian.

7.2 Model limitations

There are clear limitations connected with the use of a closed, commercial model such as ChatGPT. There is limited knowledge concerning its training data and the model weights are not shared openly. This means that there is a certain possibility of data leakage and there are also no possibilities for further fine-tuning of the model and subsequent evaluation. Unfortunately there are currently no freely available Norwegian instruction-tuned generative models, however, in future efforts we do aim to experiment with other openly available multilingual models. As mentioned above, since the ELSA dataset has only recently been released, there is no possibility for data leakage of the specific labels annotated there.

Acknowledgements

The work documented in this publication has been carried out within the NorwAI Centre for Research-based Innovation, funded by the Research Council of Norway (RCN), with grant number 309834.
We would like to thank the anonymous reviewers for their helpful comments, and Håkon Liltved Hyrve for research assistance.

References

Meysam Alizadeh, Maël Kubli, Zeynab Samei, Shirin Dehghani, Juan Diego Bermeo, Maria Korobeynikova, and Fabrizio Gilardi. 2023. Open-source large language models outperform crowd workers and approach chatgpt in text-annotation tasks. *ArXiv*, abs/2307.02179.

Jeremy Barnes, Lilja Øvrelid, and Erik Velldal. 2021. If you've got it, flaunt it: Making the most of fine-grained sentiment annotations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 49–62, Online. Association for Computational Linguistics.

Victoria Bobicev and Marina Sokolova. 2017. Inter-annotator agreement in sentiment analysis: machine learning perspective. In *International Conference Recent Advances in Natural Language Processing*, pages 97–102.

Evan Dufraisse, Adrian Popescu, Julien Tourille, Armelle Brun, and Jerome Deshayes. 2023. MAD-

TSC: A multilingual aligned news dataset for target-dependent sentiment classification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8286–8305, Toronto, Canada. Association for Computational Linguistics.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences of the United States of America*, 120.

Evgeny Kim and Roman Klinger. 2018. Who feels what and why? annotation of a literature corpus with semantic roles of emotions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1345–1359, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Lilja Øvrelid, Petter Mæhlum, Jeremy Barnes, and Erik Velldal. 2020. A fine-grained sentiment dataset for Norwegian. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5025–5033, Marseille, France. European Language Resources Association.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Simon Provoost, Jeroen Ruwaard, Ward van Breda, Heleen Riper, and Tibor Bosse. 2019. Validating automated sentiment analysis of online cognitive behavioral therapy patient texts: An exploratory study. *Frontiers in Psychology*, 10.

Egil Rønningstad, Erik Velldal, and Lilja Øvrelid. 2022. Entity-level sentiment analysis (ELSA): An exploratory task survey. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6773–6783, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Jakub Šmíd and Pavel Přibáň. 2023. Prompt-based approach for Czech sentiment analysis. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 1110–1120, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Erik Velldal, Lilja Øvrelid, Eivind Alexander Bergem, Cathrine Stadsnes, Samia Touileb, and Fredrik Jørgensen. 2018. NoReC: The Norwegian review corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

A Experiment gpt0802

The following are the Assistant instructions and the prompt used for retrieving a replay containing a JSON-formatted segment with the sentiment labels for each pre-identified entity in one text. All files referenced in the system instructions and the prompts, were uploaded with the parameter `purpose="assistants"`.

A.1 Assistant instructions

```
instructions=f"You are a helpful assistant designed to output sentiment classification labels. All questions are about entity-wise sentiment analysis on Norwegian texts. You will analyze the sentiment regarding one volitional entity at the time, inspecting a Norwegian text that is provided as the introduction. The reply should contain the analysis of the sentiment towards the entity submitted in the first reply, and a second reply should contain the sentiment label only, chosen from this list: ‘[’Positive-Standard’, ’Positive-Slight’, ’Neutral’, ’Negative-Slight’, ’Negative-Standard’ ]‘. ’Neutral’ is the most common label. ’Positive-Slight’ and ’Negative-Slight’ are used if an entity receives slight, vague or uncertain sentiment. Otherwise, the ’Positive-Standard’ and ’Negative-Standard’ labels are used for all clear sentiments expressed towards the entity. You should not refer to common knowledge about an entity, but strictly analyze the sentiment conveyed in the given text. If both positive or negative sentiments exist, you must decide what is the prevalent or overall strongest sentiment conveyed in the text regarding the enity in question. You should make use of the instructions in the file {instructions.id} for determining the sentiment and give a json with the entities and their corresponding sentiments. In {pretraining.id} you will find some gold examples of this analysis."
```

A.2 Example prompt

The prompts were given in Norwegian. The following is a simple translation into English for one example text and the relevant entities in the document:

We are going to analyze the entities in the document ’file-OtlWmi9LJgyOMsB3dKelNDZK’. The text mentions these 8 entities: [Jamie, Jared Fraser, Claire, Caitriona Balfe, Diana Gabaldons, Black Jack Randalls, Ludvig XV of France, Sam

Heughan]. Your task is to assign a sentiment label that the text in file-OtlWmi9LJgyOMsB3dKelNDZK communicates regarding each entity, according to the system instructions for the assistant.