Good or Bad News? Exploring GPT-4 for Sentiment Analysis for Faroese on a Public News Corpora

Iben Nyholm Debess[1], Annika Simonsen[2], Hafsteinn Einarsson[2]
University of the Faroe Islands[1], University of Iceland[2]
V.U. Hammershaimbs gøta 22, 100 Tórshavn, The Faroe Islands[1]
Sæmundargata 2, 102 Reykjavík, Iceland[2]
ibennd@setur.fo, annika@hi.is, hafsteinne@hi.is

Abstract

Sentiment analysis in low-resource languages presents unique challenges that Large Language Models may help address. This study explores the efficacy of GPT-4 for sentiment analysis on Faroese news texts, an uncharted task for this language. On the basis of guidelines presented, the sentiment analysis was performed with a multi-class approach at the sentence and document level with 225 sentences analysed in 170 articles. When comparing GPT-4 to human annotators, we observe that GPT-4 performs remarkably well. We explored two prompt configurations and observed a benefit from having clear instructions for the sentiment analysis task, but no benefit from translating the articles to English before the sentiment analysis task. Our results indicate that GPT-4 can be considered as a valuable tool for generating Faroese test data. Furthermore, our investigation reveals the intricacy of news sentiment. This motivates a more nuanced approach going forward, and we suggest a multi-label approach for future research in this domain. We further explored the efficacy of GPT-4 in topic classification on news texts and observed more negative sentiments expressed in international than national news. Overall, this work demonstrates GPT-4's proficiency on a novel task and its utility for augmenting resources in low-data languages.

Keywords: Sentiment Analysis, Faroese, News text, GPT-4

1. Introduction

News promise to inform readers of good or bad events, but can Large Language Models (LLMs) discern sentiment as reliably as human readers for low-resourced languages with few speakers? We investigate this question through a case study of sentiment analysis for Faroese news. Faroese, an Insular Scandinavian language spoken in the Faroe Islands (population 55,000[1]), provides a good example of a low-resourced language. While sentiment analysis has become a pivotal tool in natural language processing, most research has focused on major languages and emotionally laden text genres like social media and reviews. However, for low-resource languages such data can be difficult to find. Prior work in sentiment analysis for low-resource languages, makes use of such resources, for example, twitter data, when available (Muhammad et al., 2022), or look to alternative data sources such as news texts (Stefanovitch et al., 2022; Kolb et al., 2022) or even literary works (Pavlopoulos et al., 2022). BERT-based models are then fine-tuned on labelled sentiment analysis data. However, for severely under-resourced languages, such as Faroese, a sizeable labelled sentiment analysis dataset for training a BERT model is often not available. Having a capable LLM, like GPT-4, opens the possibility of analysing existing text to bootstrap such a language resource creation process.

The most commonly available texts in Faroese are news articles, which pose unique challenges for sentiment analysis. News strive for an objective tone, yet the stories covered can inherently contain positive or negative sentiment. With the rise of LLMs like GPT-4, the question arises whether we now have the ability to analyse sentiment in diverse linguistic settings? GPT-3.5 and GPT-4 have shown a stunning ability to solve tasks in English (Lopez-Lira and Tang, 2023; Zhang et al., 2023; Wang et al., 2023), but there is limited published evidence of broad capabilities of LLMs for low-resourced languages.

In this paper, we investigate whether GPT-4 can reliably classify sentiment in Faroese news texts, compared to human annotators. Through a new dataset of Faroese news, hand-labelled for sentiment at the sentence and document level, we probe the limits of cross-linguistic transfer for this task. Can advanced neural models discern positivity, negativity, and neutrality within low-resource texts? Our study provides an important test case for sentiment analysis in an understudied domain. We also delve into how effectively GPT-4 can categorise topics and gauge sentiment in Faroese news articles on both national and international topics.

The results show promise in sentiment analysis performance by GPT-4 on Faroese news text. The

[1] https://hagstova.fo/fo/folk/folkatal/folkatal

LREC-COLING 2024, pages 7814–7824
20-25 May, 2024.

human inter-annotator agreement was moderate at the sentence level and substantial at the document level. Comparing agreement with GPT-4, the LLM demonstrates reliable performance. Analysing the results, the complexity of news texts as a domain becomes evident, which is why we suggest a multi-label approach going forward in the discussion section. The dataset is available online[2].

2. Literature review

2.1. Sentiment analysis

Sentiment analysis, also referred to as opinion mining, is the computational treatment of opinion, sentiment, and subjectivity in text (Pang and Lee, 2008; Liu, 2020). Sentiment analysis can be used for a wide array of applications, such as predicting how the stock market will behave (Bollen et al., 2011), help companies analyse reviews of their products or even revealing a global positivity bias in natural human languages (Dodds et al., 2015). Analysing sentiment is no trivial task, especially when the text in question is news. Sentiment analysis generally handles subjective texts, such as tweets, blogs, or reviews, where authors express their opinions freely, which can be contrasted with the objective tone intended in news. However, news articles are not immune to the infusion of lexically expressed opinions, albeit at a lesser frequency compared to reviews (Balahur et al., 2013). Even as news authors strive to remain objective, sentiment is often expressed implicitly in the text through narrative framing or selective presentation of facts. Consequently, readers find themselves navigating not only explicit information but also the undertones of sentiment, which can significantly shape their perception of the narrated events.

2.2. Approaches

Examining news content closely reveals three distinct perspectives: that of the author, the reader, and the text itself (Balahur et al., 2013). Although authors are expected to maintain neutrality, their individual stance might inadvertently be disclosed through their writing style. Conversely, readers interpret news through a lens shaped by personal factors such as cultural background and education. The text itself, meanwhile, has the potential to manifest sentiments either overtly or covertly (e.g., depending on world knowledge). Balahur et al. (2013) therefore advocated for a text-centric view, as this approach does not necessitate inferring the intended meaning. However, they emphasised that these three perspectives should be addressed differently, suggesting that some of them might more aptly belong to studies on perspective determination or news bias research.

Recent approaches to sentiment analysis have suggested a reader-centric focus, aiming to gauge the average reader's response to news narratives. This was notably applied in the creation of a Slovene web-crawled news corpus annotated with sentiment (Bučar et al., 2018). To avoid having their data affected by the six annotator's personal opinions, Bučar et al. (2018) instructed annotators to annotate from the perspective of "an average Slovene web user". Furthermore, the annotators simply had to annotate how reading the news made them feel. If a sentence contained more than one sentiment, the annotator then had to choose the most dominant sentiment. If this was not possible, the annotator was to choose neutral. Similarly, Van Hee et al. (2021)'s guidelines for the Dutch-language news article corpus instructs annotators "that the annotations should be made from a European/Western viewpoint". This approach is echoed in guidelines proposed by Mukta et al. (2021) for sentiment analysis in Bengali, where annotators are asked to annotate from "the point of view in which most people would agree with".

The dynamic landscape of sentiment analysis has recently witnessed Engelund (2023) delving into the formulation of guidelines for analysing Danish news articles. This study, inspired by previous works, accentuates the necessity for a defined framework to approach sentiment analysis in news, contemplating a reader-centric analysis that respects the multicultural and diverse perspectives of the modern world. Engelund (2023)'s guidelines therefore served as a starting point for the guidelines that we developed for our study.

2.3. Related works

2.3.1. Low-resource Sentiment Analysis using LMs

Recent work on Sentiment Analysis for low-resource languages has revolved around different approaches and different text domains - but a common factor has been the incorporation of a language model, such as BERT. A model like BERT will generally require fine-tuning on a few thousand examples to perform well. If a low-resource language has enough speakers, it is usually possible to obtain Twitter data to work with. This method was used by Muhammad et al. (2022), who were able to create the first large-scale human-annotated Twitter sentiment dataset for the four languages in Nigeria (Hausa, Igbo, Nigerian-Pidgin, and Yorùbá). Using this dataset, they created a sentiment lexicon and evaluated a range of pre-trained models, mBERT and AfriB-

[2] https://huggingface.co/datasets/hafsteinn/faroese_sentiment_analysis

ERT, and transfer strategies on the dataset. Their findings suggest that language-specific models and language-adaptive fine-tuning resulted in the best performance.

There has also been work done on low-resource languages using news text instead of Twitter as data; Stefanovitch et al. (2022) released the first ever publicly available annotated dataset for sentiment classification and semantic polarity dictionary for Georgian in 2022, which is based on news articles from the Europe Media Monitor. Their best performing model was XLM-Roberta, a transformer-based model that was trained on a version of the Georgian corpus translated into English (although, this was possibly due to overfitting). Pavlopoulos et al. (2022) also used translated data for their sentiment annotation task of the first Book of Iliad; the text from the first Book of Iliad was translated into modern Greek, before annotators labelled the sentiment verse by verse. They experimented with a pre-trained model, GreekBERT, and fine-tuned it to estimate the sentiment of the data, resulting in a low error rate (Pavlopoulos et al., 2022).

While most research on sentiment analysis within news coverage adopts a broad approach, there are studies that focus on specific facets of sentiment. One such study, conducted by Kolb et al. (2022), centred on Austrian German, a low-resource language, and chose to hone in on a particular domain by developing the Austrian German sentiment dictionary, ALPIN. This dictionary draws upon Austrian news media in the political sphere, an Austriacism list, and a posting dataset derived from a well-known Austrian news outlet. Notably, the study limited its sentiment analysis to sentences featuring the names of Viennese politicians. Instead of employing language modelling, the research applied the SPLM algorithm, supplemented by crowd-sourcing and Best-Worst-Scaling techniques. Other studies, focusing on Danish, Norwegian, Dutch, and English, have similarly adopted a narrow lens, utilising embedding models in their approach (de Vries, 2022).

In the case of other Scandinavian languages, the emphasis has predominantly been on the development of sentiment lexicons (for Danish, see Nielsen (2011), Lauridsen et al. (2019) and Nimb et al. (2022), for Swedish, see Rouces et al. (2018), for Norwegian, see Øvrelid et al. (2020)) and the use of said sentiment lexicons (Enevoldsen and Hansen, 2017; Liu et al., 2019; Schumacher et al., 2019; Borg and Boldt, 2020). To our knowledge, no work has been published on Sentiment Analysis for Icelandic, the language closest related to Faroese.

2.3.2. Sentiment Analysis and ChatGPT

Much of the emerging work on Sentiment Analysis, specifically for English-centred contexts, makes use of LLMs such as ChatGPT. Recent research underscores ChatGPT's capacity to transcend conventional analytical approaches. Lopez-Lira and Tang (2023) demonstrate ChatGPT's potential to predict stock market returns by using sentiment analysis of news headlines and found that ChatGPT outperformed traditional analysis methods. Conversely, Qin et al. (2023), found that text-davinci-003 (GPT-3.5) is slightly superior to ChatGPT-3.5 with regards to sentiment analysis applied to the SST2 database (Socher et al., 2013). More recently, Zhang et al. (2023) evaluated LLMs' performance of sentiment analysis across 13 tasks on 26 datasets and compared the results against small language models (SLMs) trained on domain-specific datasets. Their findings endorse LLMs for few-shot learning endeavours in scenarios with constrained annotation assets. While LLMs thrive in straightforward sentiment analysis assignments, their proficiency diminishes in tasks demanding a deep-seated understanding and nuanced sentiment analysis (Zhang et al., 2023).

For low-resource languages, the situation seems to be more complicated; Bang et al. (2023) find that ChatGPT-3.5's ability to perform a sentiment analysis task on a low-resource language (such as Javanese) can be good, but ChatGPT-3.5 still has a very limited understanding of extremely low-resource languages (such as Buginese), suggesting that future work should focus on enhancing LLMs ability to understand sentiment in extremely low-resource languages. Hasan et al. (2023) created a large, manually annotated dataset in Bangla, a low-resource language, consisting of news, tweets and Facebook comments and used it to investigate the sentiment analysis performance of several language models, including Flan-T5, GPT-4, and Bloomz, in zero- and few-shot in-context learning compared to fine-tuned models. Their research indicated that monolingual transformer-based models consistently exceeded the performance of other models, including in zero and few-shot contexts. In principle, we have a similar approach to the Bangla study, but we extract information from the language model in a structured manner. Extracting information in a structured manner is an important step for applying LLMs to annotate data in a straightforward manner.

In summation, while strides have been made in sentiment analysis through LLMs such as ChatGPT, a conspicuous gap remains, especially concerning low-resource languages. However, latest studies indicating good performance of LLMs

in sentiment analysis of low-resource languages (Hasan et al., 2023; Zhang et al., 2023) motivate the setup of our experiment.

3. Methodology

This experiment was broken down into the following stages; collecting text for annotating, defining a prompt for GPT-4's sentiment analysis, having GPT-4 assign sentiment to the text on a sentence level and document level, having GPT-4 assign topic to all articles from a given topic list, extracting an equal sample for two human annotators to annotate blindly, developing annotation guidelines, re-running GPT-4 on said guidelines and comparing results.

3.1. Data source

The collected text was selected from the Basic Language Resource Kit 1.0 for Faroese (Debess et al., 2022) open source text corpus, described in Simonsen et al. (2022). The documents are short news articles published on the Faroese online news sites, *Portalurin*[3] and *Dimmalætting*[4], 44042 words in 170 articles. We created the dataset by randomly selecting 1-3 full sentences for annotation and analysis. This approach allowed for analysis on two levels (the sentence level and the document level) with a total of 225 sentences selected.

3.2. Analysis

3.2.1. Sentiment analysis

The articles were analysed both on sentence level (1-3 randomly selected sentences) and on document level (full article). The sentiment analysis was multi-class: positive (1), neutral (0), and negative (-1). Every sentence and every article were given one score each.

3.2.2. Topic analysis

For each news article, GPT was instructed to assign topics to it from a pre-selected list (shown in Figure 4) in a multi-label manner, i.e., each article could have more than a single topic. GPT-4 assigned topics for every article and human annotators verified or corrected. Additionally, human annotators labelled every article to be international, national or mixed.

3.3. Human annotators and guidelines

The human annotators were two linguists, both native speakers of Faroese. The linguists had to define a set of annotation guidelines during the annotation process. There are no official guidelines for annotating news paper text, but the linguists chose to use guidelines from a recent Danish master thesis (Engelund, 2023) on how to define guidelines for news paper articles as a baseline. Annotation guidelines will always need to be adjusted to fit whatever data it is used on, so having a baseline gave the annotators a starting point. Two initial adjustments were made:

- We use full articles, not just title and headline.
- When the article gets too technical and complicated, we lean towards neutral.

Using Engelund (2023)'s adjusted annotation guidelines, the linguists annotated the first 30 texts (sentence level and document level) individually and then compared annotations to see where the disagreement was. Then after defining the guidelines further, the linguists annotated the 30 next sentences and compared again, before defining final guidelines and annotating the rest of the sentences. After the final revision of guidelines was made, the first 60 sentences were annotated again to ensure consistency. During this process, the linguists also checked if GPT-4's assigned topic was relevant or not.
The final guidelines were as follows:

- Annotate the sentence or document based on the overall sentiment that it invokes (positive, neutral or negative).
- Judge the sentence or document from the perspective of the average Faroese news reader.
- Take local knowledge and culture into consideration, but do not include personal opinions of politics or religion etc.
- Don't include the author's personal opinion - only care about the sentiment that is invoked in the reader's mind.
- Take textual context into consideration when annotating.
- If a sentence or document is both equally positive and negative, the annotation is assigned as neutral.
- If it is not possible to estimate sentiment due to the political subtext of the sentence or document (e.g. "Donald Trump won the election"), then it is assigned as neutral.

These guidelines align with the ones from previous work on news sentiment (Bučar et al., 2018; Van Hee et al., 2021; Mukta et al., 2021), both regarding reader-centricity, including societal context, having an 'average' viewpoint, and choosing the dominant sentiment, when possible.

[3] https://portal.fo/
[4] https://www.dimma.fo/

3.4. Prompting approach

We prompt GPT-4 with a temperature of 0 using their function API to extract information in a structured manner. For each news article, the output needs to respect a given JSON schema as shown in Listing 1. In the schema, we implicitly ask GPT-4 to split the text into sentences. Each sentence is translated, assigned a sentiment, the news article is assigned one or more topics, and an overall sentiment for the news article is assigned. We also study a variation where we leave out the translation requirement.

It should be noted that GPT-4 sometimes failed in splitting the sentences correctly, and such splits where left out. Furthermore, in a few cases it suggested topics that were not in the list of allowed values, which was fixed by the annotators.

```
"name": "sentiment_analysis",
"description": "The function analyses text that has been translated from
Faroese to English. The input translation should be of exceptionally
high quality.",
"parameters": {
    "type": "object",
    "properties": {
        "sentence_analysis_list": {
            "type": "array",
            "items": {
                "type": "object",
                "properties": {
                    "original": {"type": "string"},
                    "translation": {"type": "string"},
                    "sentiment": {
                        "type": "integer",
                        "description": "A number representing the overall
                        sentiment of the sentence. The number is -1, 0 or
                        1, where -1 is negative, 0 is neutral, and 1 is
                        positive.",
                    }
                }
            }
        },
        "topic": {
            "type": "array",
            "items": {
                "type": "string",
                "enum": ["Local news", "International
                news", "Culture", "Sports", "Science",
                "Economy", "Education", "Fishing",
                "Technology", "Entertainment", "Politics",
                "Health", "Crime and Justice", "Event
                announcement", "Opinion piece", "Other"]
            }
        },
        "overall_sentiment": {
            "type": "integer",
            "description": "A number representing the overall sentiment
            of the text. The number is -1, 0 or 1, where -1 is negative,
            0 is neutral, and 1 is positive.",
        }
    },
    "required": ["sentence_analysis_list","topic","overall_sentiment"]
}
```

Listing 1: JSON schema for extracting sentiment at the sentence level and the overall sentiment of the text.

Additionally, we study a few-shot instruction for sentiment analysis reflecting the linguists' guidelines that we either include or leave out. The instructions are shown in italics below[5].

Sentiment analysis refers to annotating the dominant sentiment invoked in the average Faroese reader, when reading a given text or sentence.
The sentiment analysis should include knowledge of societal and textual context. The sentiment analysis should leave out political and personal opinions. When positive and negative sentiment is equally present, analyse as being neutral.
Examples:
Example of a sentence invoking a negative sentiment, represented by the number -1: "A citizen called the police several times last night to complain about lack of sleep caused by noise."
Example of a sentence invoking a neutral sentiment, represented by the number 0: "Johannes Absalonsen represents the committee at the event."
Example of a sentence invoking a positive sentiment, represented by the number 1: "Here we are supported and encouraged, and here we join in celebrating big and small victories."

[5]For the purpose of this article, examples are written in English. Original prompt examples were Faroese.

3.5. Performance metrics

Cohen's Kappa was calculated for inter-annotator agreement between the human annotators and between the human annotators and GPT-4 (Cohen, 1960).

4. Results

4.1. Inter-Annotator Agreement and Consensus Formation

At the sentence-level, the human inter-annotator agreement was moderate with a Cohen's Kappa of 0.59. Annotator agreement between humans and GPT-4 with different prompt approaches ranged from 0.47 to 0.58 (see Table 1). We studied two ways to form a consensus between the human annotators. First, a strict agreement approach where we disregarded all examples where annotators A and B disagreed. After reviewing the confusion matrices of the annotators we defined another consensus approach. We saw that the annotators were more conservative in assigning positive and negative sentiments than GPT because the confusion was mostly on the centre row of the confusion matrix (see Figure 1). As a result, we defined a relaxed agreement consensus where a sentence was labelled as positive if at least one annotator marked it as positive and the other marked it as positive or neutral. Similarly, we assigned a negative label if one annotator marked it as negative and the other as negative or neutral. A sentence was considered neutral if both annotators assigned it a neutral label. This brought the agreement up to 0.70, which is indicative of substantial agreement. Confusion matrices for the relaxed agreement consensus label and the different prompting configurations is shown in Figure 2.

At the document level, the human inter-annotator agreement was substantial with a Cohen's kappa value of 0.65. Annotator agreement between humans and GPT-4 with different prompting configurations ranged from 0.43 to 0.53 (see Table 1). Similarly to the sentence level agreement, we see greater agreement with the relaxed agreement consensus approach, but the improvement

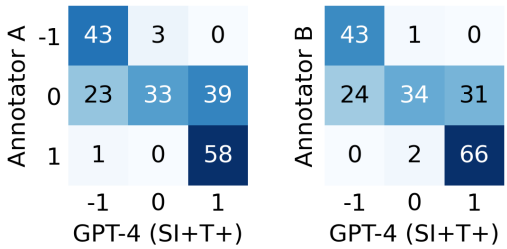

Figure 1: Confusion matrices for sentiment on sentences for the two annotators and the best performing prompting setup with translation and sentiment instructions.

		SI+T+	SI+T-	SI-T+	SI-T-
Sentence	A	0.52	0.50	0.49	0.47
	B	0.58	0.50	0.53	0.55
	SA	0.52	0.50	0.49	0.47
	RA	0.70	0.61	0.64	0.64
Document	A	0.52	0.53	0.44	0.45
	B	0.43	0.47	0.44	0.45
	SA	0.52	0.53	0.44	0.45
	RA	0.55	0.57	0.51	0.52

Table 1: Cohen's Kappa values between annotators (A = Annotator A, B = Annotator B, SA = Strict agreement, RA = Relaxed agreement) and prompting configurations. Shorthand notations: SI = Sentiment Instructions, T = Translation and + means the feature was used in the prompt and - that it was not.

is not as large. Confusion matrices for the relaxed agreement consensus label and the different prompting configurations is shown in Figure 3.

4.2. Topic Analysis

The annotators reviewed the topics assigned by GPT-4 and corrected mistakes in 15 out of 170 news articles, i.e., the error rate was quite low at 8.8%. The corrected topic distribution is shown in Figure 4.

The annotators labelled each article as national, international or mixed resulting in 181, 54 and 10 articles from each category, respectively. We compare the sentiments assigned to the National and International categories in Figure 5 using the different annotation methods. We observe that national news have proportionally more articles with a positive sentiment. At the document level, the agreement with the relaxed consensus ranges from 0.38 (SI-T-) to 0.57 (SI+T-) for international news and from 0.47 (SI-,T+) to 0.53 (SI+T+) for national news.

5. Discussion

5.1. Inter-annotator agreement

The results show moderate inter-annotator agreement between human annotators and GPT-4, indicating that GPT-4 can classify sentiment in Faroese news texts quite reliably. Forming a relaxed agreement consensus improved the agreement score overall, and specific prompting configurations indicated substantial agreement. Given the low-resource setting, these results are promising and prove the possibility of automated data labelling for Faroese via GPT-4. With no baseline or previous study of this specific task, though, future studies will need to reveal what is the best automated method to approach this task.

When analysing the agreement results with regard to the different prompt configurations, we see that the approach requiring translation in GPT-4's labelling process does not improve agreement overall. We see a small increase in agreement on sentence level, and conversely we see a small decrease on document level. This is surprising as we had hypothesised that translating the Faroese sentences into English and analysing the sentiment of the English sentences would make the analysis more accurate, as GPT-4 generally performs better on English than small languages (Chang et al., 2023). Furthermore, a quick informal review revealed good translation quality. However, we do see an increase in agreement when configuring the prompt to include the few-shot sentiment instruction. This was expected, as the instructions were formed to align with human annotator guidelines and GPT models have been shown to benefit from few-shot prompting.

5.2. Topics

GPT-4 performed well in the topic annotation task with few errors. The majority of the articles were about Sports, Local news, or Health. The amount of articles about Health is probably due to the time period in which the articles were written, as COVID-19 was spreading at that time. The text about pandemic issues were especially difficult to annotate. One could be inclined to label every sentence about COVID-19 as negative by default, but what should one do about a sentence such as *"Due to the rise of the new COVID-19 variant, Israel closes its border"*. For some readers, the sentiment invoked is negative, because it is news about COVID-19, but for others it might invoke positive sentiment, because Israel is taking precautions to stop the spread of the variant.

The labelling of **national**, **international** or **mixed** news articles was motivated by the hypothesis that GPT-4 would perform better in annotating sentiment when topics were international, as these topics should be more familiar to the model than ob-

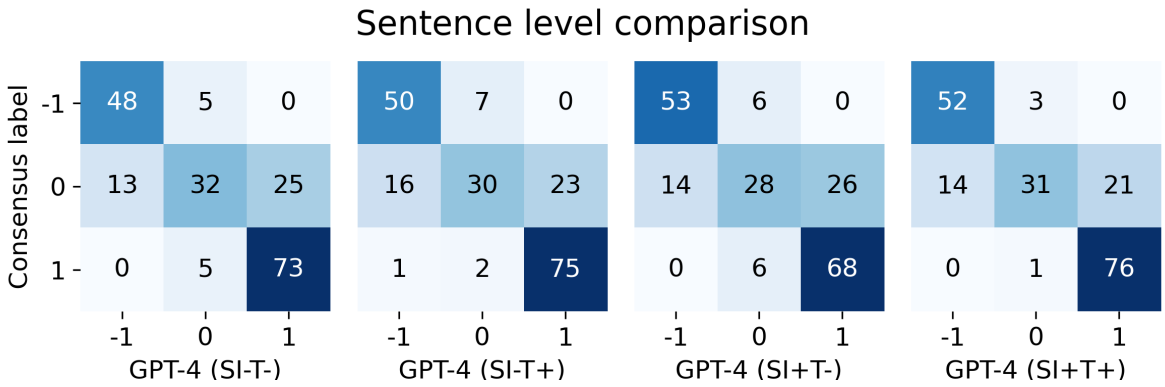

Figure 2: Confusion matrices for sentiment on sentences between the relaxed agreement consensus approach and the different prompting configurations.

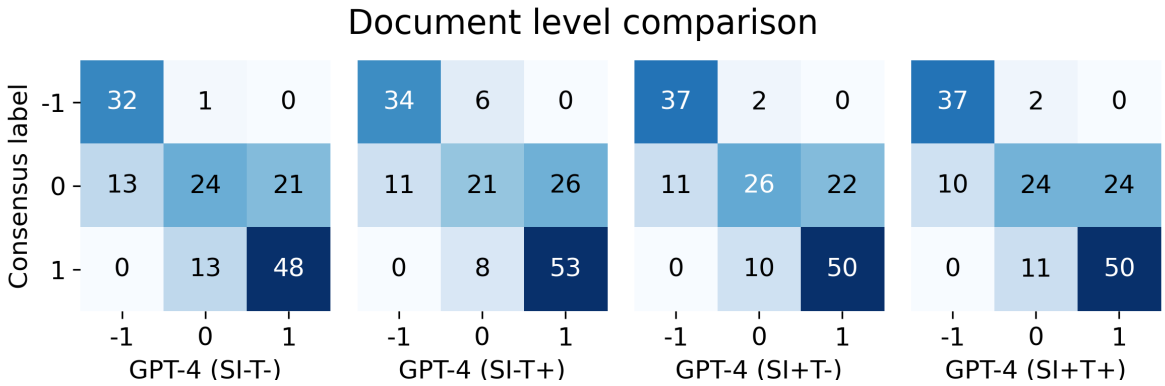

Figure 3: Confusion matrices for sentiment on documents between the relaxed agreement consensus approach and the different prompting configurations.

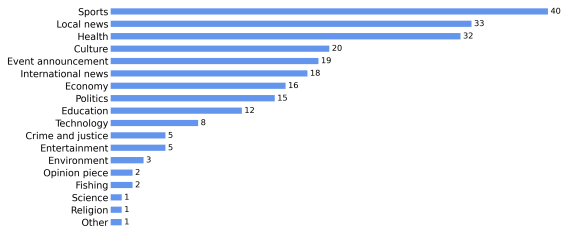

Figure 4: Topic distribution of the news articles.

scure local news. Surprisingly, we did not observe large differences in agreement between national and international news. However, we did see GPT-4 being more polar (especially negative) about the international news than the human annotators, but this was also the case for national news, just to a lesser extent. Looking at the overall analysis, local news are annotated more positively than international news, which might indicate a selection at the editorial level for bad news in the context of international news. However, a larger study over a longer time period (which excludes COVID-19) would need to confirm that hypothesis.

5.3. The domain of news text

In previous work, sentiment analysis has mostly been conducted on either sentences or shorter text, e.g. reviews, tweets, exclusively titles and headlines (Engelund, 2023) or specific sentences filtered out from datasets (de Vries, 2022; Kolb et al., 2022). In our experiments, we not only annotated sentences but also the full articles that they appeared in (document level). It proves difficult to assign overall sentiment to a full news article, especially the longer ones. A long news article will often include both positive and negative sentiment, so in the case of our dataset and the guidelines followed, they often end up being annotated as neutral.

As a part of our methodology, random sentences were chosen from the news articles. However, that process might require reconsideration when evaluating news. The stronger sentiments, positive and negative, tend to be concentrated at the beginning of an article (Bučar et al., 2018). This tendency could be considered a justification for using sentences in the same place of the document (either beginning, middle or end) for all texts, rather than randomly selecting sentences from all parts of an article.

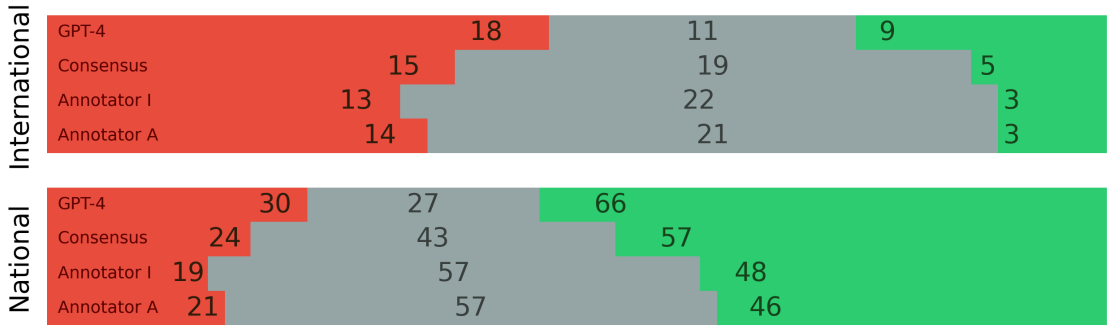

Figure 5: Sentiment distribution for international and national news using different annotation approaches. The consensus corresponds to the relaxed agreement consensus approach and the GPT-4 annotator is the one with sentiment instructions and translation.

The moderate to substantial inter-annotator agreement can be seen as indicative of two issues: 1) the complex nature of news text, and 2) the need for clearer annotator guidelines. These two issues relate to each other as the need for clearer guidelines stems from the textual domain of news being too sentimentally complex, especially in document level analysis. The articles are often ambiguous and can invoke multiple sentiments in one reader, as well as being sensitive to invoking different sentiments based on different readers' perspective.

The example *"Today's poll revealed that Sweden will have its first female Head of Parliament."* is sentimentally ambiguous, as it can invoke either positive or negative sentiment depending on reader demographics, and the political issue at hand makes it hard for the annotator to position the 'average reader'.

To address the issue of news text complexity, drawing from our project experience, we propose a **multi-label** annotation scheme as opposed to the multi-class approach we took here. The multi-label approach has demonstrated effectiveness and appropriateness across various contexts, notably in the news sector. This is exemplified by the study of Almeida et al. (2018), which concentrated on headlines and articles from newspapers, and extends to additional fields, such as the sentiment analysis of texts from social media platforms and reviews as discussed by Huang et al. (2013) and Tao and Fang (2020). This latter research encompassed various reviews. A multi-label approach enables annotation with multiple sentiments for one text, thus not forcing the annotator to estimate the weight of different sentiments in a text, a process susceptible to subjectivity. This can be contrasted with the neutral category in this work which can include strongly polarised sentiments (cf., guidelines). As an alternative, Kenyon-Dean et al. (2018) suggest adding a 'complicated' class to the annotation. Going for multi-labelling would encompass and expand this suggestion, as any text annotated with multiple sentiment labels would be 'complicated' in nature as well as include details on the complexity of the sentiment.

Additionally, we set out the advantages of **crowd-sourcing** for news sentiment analysis. Not only is this a method volume efficient for small languages, it also fits well for reader-based sentiment, which is recommended when analysing news (Bučar et al., 2018; Engelund, 2023). For the crowd-sourcing task, we plan to adjust the guidelines to ensure that annotators provide insights based on their individual perspectives, rather than attempting to represent the collective opinion of the average Faroese news reader. Crowd-sourcing addresses the annotator bias problem where the views or understanding of one or few annotators can have a large effect on the resulting annotated dataset. Ensuring a diverse group of annotators will enable us to gather data that more accurately reflects the average news reader from the Faroese population. For this particular task, it must be highlighted that labelling data from the point of view of an undefined 'average reader', while taking into account a broad range of societal, cultural and political values and still staying objective is a complex task for any annotator. Well-balanced crowd-sourcing, on the other hand, would provide sentiment analyses from the actual average reader, representing true variation in society and culture with diverse viewpoints.

6. Conclusion

The results of this pilot study agree with the findings of Hasan et al. (2023) and Zhang et al. (2023), and suggest that GPT-4 has the ability to perform sentiment annotations on Faroese, a low-resourced language. We further observe good performance in the topic classification task.

This performance highlights how LLMs such as ChatGPT can facilitate the creation of annotated datasets for languages such as Faroese, which will have positive implications for the development of Faroese language technology.

Our study confirms that analysing sentiment in news texts is indeed a difficult task and needs other approaches than sentiment analysis of other linguistic domains. Due to the difficulty of the annotation task, we consider the main limitation of this study to be the multi-class annotation process. We recommend multi-label annotation as the future path for this task.

However, we must also acknowledge that OpenAI's non-compete requirements for commercial-use of GPT-4's output limit the future applications of this dataset. Still, it can be used for evaluating multilingual sentiment classification models and could as a result spark interest in Faroese NLP.

7. Acknowledgements

AS was supported by the European Commission under grant agreement no. 101135671.

8. Bibliographical References

Almeida, A. M. G., Cerri, R., Paraiso, E. C., Mantovani, R. G., and Junior, S. B. 2018. Applying multi-label techniques in emotion identification of short texts. *Neurocomputing*, 320:35–46.

Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., Van Der Goot, E., Halkia, M., Pouliquen, B., and Belyaeva, J. 2013. Sentiment analysis in the news. pages 2216–2220, Valletta, Malta. European Language Resources Association (ELRA).

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., Do, Q. V., Xu, Y., and Fung, P. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity.

Bollen, J., Mao, H., and Zeng, X. 2011. Twitter mood predicts the stock market. *Journal of computational science*, 2(1):1–8.

Borg, A. and Boldt, M. 2020. Using VADER sentiment and SVM for predicting customer response sentiment. *Expert Systems with Applications*, 162:113746.

Bučar, J., Žnidaršič, M., and Povh, J. 2018. Annotated news corpora and a lexicon for sentiment analysis in Slovene. *Language Resources and Evaluation*, 52:895–919.

Chang, Y., Wang, X., Wang, J., Wu, Y., Zhu, K., Chen, H., Yang, L., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., and Xie, X. 2023. A survey on evaluation of large language models.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

de Vries, E. 2022. The Sentiment is in the Details: A language-agnostic approach to dictionary expansion and sentence-level sentiment analysis in news media. *Computational Communication Research*, 4(2):424–462.

Dodds, P.S., Clark, E.M., Desu, S., Frank, M.R., Reagan, A.J., Williams, J.R., Mitchell, L., Harris, K.D., Kloumann, I.M., Bagrow, J.P., Megerdoomian, K., McMahon, M.T., Tivnan, B.F., and Danforth, C.M. 2015. Human language reveals a universal positivity bias. *Proceedings of the National Academy of Sciences*, 112(8):2389–2394.

Enevoldsen, K. C. and Hansen, L. 2017. Analysing political biases in Danish newspapers using sentiment analysis. *Journal of Language Works-Sprogvidenskabeligt Studentertidsskrift*, 2(2):87–98.

Engelund, T. K. M. S. 2023. *Annotation of Sentiment Evoked by Danish News Articles: Leveraging Linguistic Corpus Annotation for Reliable Annotation and Effective Transformer Fine-Tuning*. Master of science in it & cognition thesis, University of Copenhagen, Copenhagen.

Hasan, A., Das, S., Anjum, A., Alam, F., Anjum, A., Sarker, A., and Noori, S. R. H. 2023. Zero- and Few-Shot prompting with LLMs: A comparative study with Fine-tuned models for Bangla Sentiment Analysis.

Huang, Shu, Peng, Wei, Li, Jingxuan, and Lee, Dongwon. 2013. Sentiment and topic analysis on social media: a multi-task multi-label classification approach. In *Proceedings of the 5th Annual ACM Web Science Conference*, WebSci '13, page 172–181, New York, NY, USA. Association for Computing Machinery.

Kenyon-Dean, K., Ahmed, E., Fujimoto, S., Georges-Filteau, J., Glasz, C., Kaur, B., Lalande, A., Bhanderi, S., Belfer, R., Kanagasabai, N., Sarrazingendron, R., Verma, R., and Ruths, D. 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.

Kim, S. and Hovy, E. 2004. Determining the sentiment of opinions. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 1367–1373, Geneva, Switzerland. COLING.

Kolb, T., Katharina, S., Kern, B. M. J., Neidhardt, J., Wissik, T., and Baumann, A. 2022. The ALPIN Sentiment Dictionary: Austrian language polarity in newspapers. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC'22)*, pages 4708–4716, Marseille, France. European Language Resources Association (ELRA).

Lauridsen, G. A., Dalsgaard, J. A., and Svendsen, L. K. B. 2019. SENTIDA: A new tool for sentiment analysis in Danish. *Journal of Language Works-Sprogvidenskabeligt Studentertidsskrift*, 4(1):38–53.

Liu, B. 2020. *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge university press.

Liu, P., Marco, C., and Gulla, J. A. 2019. Semi-supervised Sentiment Analysis for Under-Resourced Languages with a Sentiment Lexicon. In *INRA@ RecSys*, pages 12–17.

Lopez-Lira, A. and Tang, Y. 2023. Can ChatGPT forecast stock price movements? Return predictability and large language models.

Muhammad, S. H., Adelani, D. I., Ruder, S., Ahmad, I. S., Abdulmumin, I., Bello, B. S., Choudhury, M., C., Emezue. C., Abdullahi, S. S., Aremu, A., Jeorge, A., and Brazdil, P. 2022. NaijaSenti: A Nigerian Twitter Sentiment Corpus for Multilingual Sentiment Analysis.

Mukta, S. H., Islam, A., Khan, F. A., Hossain, A., Razik, S., Hossain, S., and Mahmud, J. 2021. A comprehensive guideline for Bengali sentiment annotation. *Transactions on Asian and Low-Resource Language Information Processing*, 21(2):1–19.

Nielsen, F. Å. 2011. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs.

Nimb, Sanni, Olsen, Sussi, Pedersen, Bolette, and Troelsgård, Thomas. 2022. A thesaurus-based sentiment lexicon for Danish: The Danish sentiment lexicon. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC'22)*, pages 2826–2832, Marseille, France. European Language Resources Association (ELRA).

Pang, B. and Lee, L. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval*, 2(1–2):1–135.

Pavlopoulos, J., Xenos, A., and Picca, D. 2022. Sentiment Analysis of Homeric text: The 1st Book of Iliad. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC'22)*, pages 7071–7077, Marseille, France. European Language Resources Association (ELRA).

Qin, C., Zhang, A., Zhang, Z., Chen, J., Yasunaga, M., and Yang, D. 2023. Is ChatGPT a general-purpose natural language processing task solver?

Rouces, J., Borin, L., Tahmasebi, N., and Eide, S. R. 2018. Defining a Gold Standard for a Swedish Sentiment Lexicon: Towards higher-yield text mining in the Digital Humanities. In *Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference*, pages 219–227, Helsinki, Finland. CEUR-WS.

Schumacher, G., Hansen, D., van der Velden, M. A., and Kunst, S. 2019. A new dataset of Dutch and Danish party congress speeches. *Research & Politics*, 6(2).

Simonsen, A., Lamhauge, S. S., Debess, I. N., and J., Henrichsen. P. 2022. Creating a Basic Language Resource Kit for Faroese. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC'22)*, pages 4637–4643, Marseille, France. European Language Resource Association (ELRA).

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Stefanovitch, N., Piskorski, J., and Kharazi, S. 2022. Resources and experiments on sentiment classification for Georgian. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC'22)*, pages 1613–1621, Marseille, France. European Language Resources Association (ELRA).

Tao, J. and Fang, X. 2020. Toward multi-label sentiment analysis: a transfer learning based approach. *Journal of Big Data*, 7(1):1.

Van Hee, C., De Clercq, O., and Hoste, V. 2021. Exploring implicit sentiment evoked by fine-grained news events. In *Proceedings of*

the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (EACL 2021), pages 138–148. Association for Computational Linguistics.

Velldal, Erik, Øvrelid, Lilja, Bergem, Eivind Alexander, Stadsnes, Cathrine, Touileb, Samia, and Jørgensen, Fredrik. 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).

Wang, Z., Xie, Q., Ding, Z., Feng, Y., and Xia, R. 2023. Is ChatGPT a good Sentiment Analyzer? a preliminary study.

Zhang, W., Deng, Y., Liu, B., Pan, S. J., and Bing, L. 2023. Sentiment Analysis in the era of Large Language Models: A reality check.

Øvrelid, L., Mæhlum, P., Barnes, J., and Velldal, E. 2020. A Fine-grained Sentiment dataset for Norwegian.

9. Language Resource References

Debess, I. N. and Lamhauge, S. S. and Simonsen, A. and Henrichsen, P. J. and Hofgaard, E. and Johannesen, U. and Hammer, P. M. J. and Brimnes, G. H. and Thomsen, E. M. D. and Poulsen, B. 2022. *Basic LAnguage Resource Kit 1.0 for Faroese*. Talutøkni. OpenSLR. [link].