# Incorporating Syntax and Lexical Knowledge to Multilingual Sentiment Classification on Large Language Models

**Hiroshi Kanayama, Yang Zhao, Ran Iwamoto, Takuya Ohko**
IBM Research
{hkana@jp., YangZhao@, ran.iwamoto1@, ohkot@jp.}ibm.com

## Abstract

This paper exploits a sentiment extractor supported by syntactic and lexical resources to enhance multilingual sentiment classification solved through the generative approach, without retraining LLMs. By adding external information of words and phrases that have positive/negative polarities, the multilingual sentiment classification error was reduced by up to 33 points, and the combination of two approaches performed best especially in high-performing pairs of LLMs and languages.

## 1 Introduction

The generative approach using Large Language Models (LLMs) performs very well even for non-generative tasks such as topic classification (Brown et al., 2020) and sentiment polarity judgment (Zhao et al., 2023). LLMs have garnered attention, particularly for their capability to generalize languages in addressing those tasks without the need for task-specific training using ground truth. Many tasks can be solved in a zero-shot manner by providing appropriate prompts. However, the language coverage highly depends on LLMs, and the accuracy can be quite low if the model received insufficient training or tuning on specific languages. Retraining LLMs requires a high cost to cover additional languages, and it tends to cause catastrophic forgetting in languages that have already been covered (Winata et al., 2023). In recent attempts at knowledge editing, difficulties such as knowledge conflicts have been pointed out (Li et al., 2023), and it is not easy to apply to many languages.

Thus, we attempt to enhance multilingual sentiment classification based on the generative approach without computationally intensive processes for retraining and tuning LLMs. As depicted in Figure 1, our approach obtains lexical and syntactic information from an external sentiment extractor and inserts it into the prompt. This approach

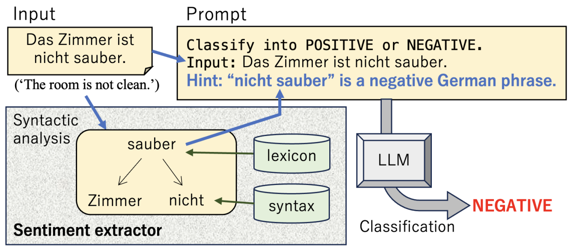


Figure 1: Concept of sentiment extraction using the prompt supported by an external sentiment extractor.

requires knowledge specific to tasks and languages but has advantages such as stability of outputs and controllability to add arbitrary knowledge to adapt to new languages or domains.

The contributions of this paper are threefold: (1) to integrate syntactic and lexical knowledge with prompts for LLMs, (2) to evaluate sentiment classification using six publicly available LLMs and parallel data in 15 languages, and (3) to demonstrate the positive impacts by integrating two components without any retraining.

## 2 Related Work

### 2.1 Multilingual Sentiment Extraction

This paper exploits a multilingual sentiment extraction method based on syntactic analysis and lexicon (Kanayama and Iwamoto, 2020). The extractor applies grammatical rules and a lexical dictionary to the common syntactic structure across languages defined by Universal Dependencies (UD) (Zeman and et al., 2017), and it extracts positive and negative phrases for many languages with high precision (although recall tends to be low for lower-resource languages), taking valence shifting such as negation into consideration.

Note that the sentiment extractor may detect one or more polar expressions from a single sentence or may not output anything when there is no explicit expression detected, and thus it does not directly solve the task of sentence-level sentiment *classification* discussed in Section 3.



*Findings of the Association for Computational Linguistics: ACL 2024*, pages 4810–4817
August 11-16, 2024 

### 2.2 Text Classification via Reasoning

Sun et al. (2023) proposed a prompting method, CARP, that enhances text classification by encouraging models to conduct reasoning. In CARP, a command to list words with polarity is added to the prompt, and it asks the model to describe the reason for classification. This strategy increased the accuracy of sentiment classification in most datasets.

In this paper, as shown in Figure 1, external knowledge is integrated into the prompt given to LLMs instead of relying solely on the knowledge embedded in LLMs. This approach enables us to broaden the language coverage and observe the characteristics of each LLM and the capability of knowledge extension using external components.

## 3 Sentiment Classification using Generative Model

This section describes the method to predict sentence polarity using an LLM and our proposal to integrate the output of an external sentiment extractor with a prompt.

### 3.1 Prompting

The task is binary classification of an input sentence into `POSITIVE` or `NEGATIVE`. Here we provide the following English prompt[1] to an LLM, in a zero-shot manner without providing labeled examples.

> Classify the next `[lang]` sentence into POSITIVE or NEGATIVE. Please just answer the label.
> input: `[input sentence]`
> output:

`[lang]` is filled with the language name such as "Arabic" and "Japanese". Then we expect the model generates either `POSITIVE` or `NEGATIVE` after "`output:`".

### 3.2 Integration with Sentiment Extractor

To enhance the LLM's decision, when the sentiment extractor detects a polar expression, we insert a hint into the prompt between the 'input' and 'output' lines[2].

> hint: "`[word]`" is a(n) `[lang]` word which has a `[polarity]` meaning.

`[word]` is the surface form of the word that was detected to have a polarity, and `[polarity]` is "positive" or "negative". For example, from a Chinese sentence "我們很失望，...", the sentiment extractor detects the negative verb "失望" ('be disappointed') through syntactic parsing and dictionary lookup, then the line

> hint: "失望" is a Chinese word which has a negative meaning.

is inserted in the prompt. If multiple polar expressions are detected, multiple lines are added. When a polar expression consists of multiple words (*e.g.* negation), a "phrase" is added as a hint. Refer to Appendix A.1 for its prompt.

## 4 Experimental Settings

### 4.1 Data Set

For evaluation of multilingual sentiment classification, we use Parallel Sentiment[3]. It consists of 106 parallel sentences (a subset of 1,000 sentences in the Parallel UD (PUD) corpora) for 19 languages annotated with positive or negative labels common for all languages, enabling us to evaluate the multilingual task with equal difficulty. We test in 15 languages that the sentiment extractor supports.

### 4.2 Language Models

We compare six LLMs that have diverse encoder-decoder architectures, language coverage in fine tuning, and model sizes.

- T5 (google/flan-t5-xxl) (Chung et al., 2024): The encoder-decoder model T5 with instruction tuning on 1,800 tasks that covers 60 languages (11 billion parameters).
- UL2 (google/flan-ul2) (Wei et al.): The encoder-decoder model UL2 with instruction tuning (20b).
- MT0 (bigscience/mt0-xxl) (Muennighoff et al., 2022): Multilingual T5 model with instruction tuning on 46 languages (13b).
- Llama2 (meta-llama/llama-2-70b-chat) (Touvron et al., 2023): Decoder model Llama tuned for chat (70b). 90 percent of the pretraining data was English.

[1]Prompts in the target language may work better, but in this paper, we always use English prompts to fairly compare the languages without concerning about the prompt quality in each language.

[2]Empirically this place is better than inserting 'hint' line before 'input' line.

[3]Distributed at `https://lrec2020.lrec-conf.org/en/shared-lrs/`.

| Model | Base | Hints | Override |
|---|---|---|---|
| No LLM | 35.4 | | **69.4** |
| T5 | 68.5 | **78.6** | **78.6** |
| UL2 | 73.9 | 75.2 | **80.2** |
| MT0 | 68.8 | 79.2 | **80.7** |
| Llama2 | 92.6 | **93.7** | 93.3 |
| Mixtral | 91.1 | **93.6** | 93.1 |
| Llama3 | 95.7 | <u>**96.6**</u> | 95.9 |

Table 1: Macro F1 scores of sentiment classification averaged in 15 languages. Bold letters show the highest score in each row and the underlined number is the highest in all cases.

- Mixtral (mistralai/mixtral-8x7b-instruct-v01) (Jiang et al., 2024): The model composed from mixture of multiple expert models and tested on multiple European languages (45b).
- Llama3 (meta-llama/llama-3-70b-instruct) (AI@Meta, 2024): The latest Llama model with instruction tuning. Though English accounts for 95 percent of the pretraining data, the bigger data and the larger vocabulary size support multilinguality (70b).

For all models, the answer is generated in a greedy decoding setting with the minimum and maximum token numbers set to 1 and 10, respectively. The MT0 model tends to output `POSITIVE` in most cases, so we adjusted the prompt as shown in Appendix A.2 to balance the prediction.

### 4.3 Sentiment Extractor

We used the sentiment extractor's outputs for the Parallel Sentiment data set obtained from Kanayama and Iwamoto (2020) on request, and distribute them as supplemental material so that the entire experiments are reproducible. The detected sentiment words and phrases are associated with detailed syntax structures in CoNLL-U format processed by StanfordNLP (Qi et al., 2018), which helped improve the sentiment extraction process even beyond the gold annotations in PUD[4].

## 5 Experimental Results

### 5.1 Overall Quality

Table 1 shows the overall results averaged across 15 languages. All numbers represent the macro F1 score of the binary sentiment classification. The weakest baseline, 35.4, corresponds to always predicting `NEGATIVE`[5] without utilizing the LLM or the sentiment extractor. A score of 69.4 was obtained by using only the sentiment extractor: answering the polarity of the sentiment expression closest to the main clause in the sentence when one or more sentiment expressions are detected[6].

Next, the six models are compared. The column labeled 'Base' shows the scores of classification using the prompt outlined in Section 3.1. The smaller models, T5, UL2, and MT0, achieve scores around 70, while the larger models show much higher scores over 90, such as 95.7 in Llama3.

The column labeled 'Hints' shows the results when the additional hints described in Section 3.2 are incorporated. By adding this information, scores increased by more than 10 points in the T5 and MT0 models, and that means 33% of errors were reduced in MT0. The absolute differences were smaller in some other models, but it is notable that 21% of errors were reduced even from the very high score in the 'Base' of Llama3.

The column labeled 'Override' reports the scores when the polarity of sentiment expression detected by the sentiment extractor replaces the LLM's decision. When multiple sentiment expressions were detected the closest one to the root node was used, and when no expression was detected the LLM's decision was used. Therefore, this score estimates the upper bound of our approach with desirable prompts to inform the external knowledge to the LLM, assuming the polarities of expressions detected by the sentiment extractor are always correct. In UL2 and MT0, 'Override' exceeds 'Hints.' This implies that the hints in the prompt were not perfectly recognized by these models. On the other hand, in the high performing models (Llama2, Mixtral and Llama3), 'Hints' achieved the highest score through the integration of LLMs and the sentiment extractor.

### 5.2 Evaluation per Language

Figures 2 to 7 show the detailed results for 15 languages. The T5 model in Figure 2 shows high 'Base' scores for Romance and Germanic lan-

[4]This is due to different annotation policies between PUD and other UD corpora used to train the parser, or the lack of lemma information in 7 out of 15 PUD corpora.

[5]`NEGATIVE` is the majority (54.7%) label in the data set, and predicting always `NEGATIVE` results in the accuracy 54.7 and the macro F1 score 35.4.

[6]When the sentiment extractor does not detect anything, the majority label `NEGATIVE` is used for prediction. It often happens, due to the design to prioritize precision over recall, and low dictionary coverage in some languages.

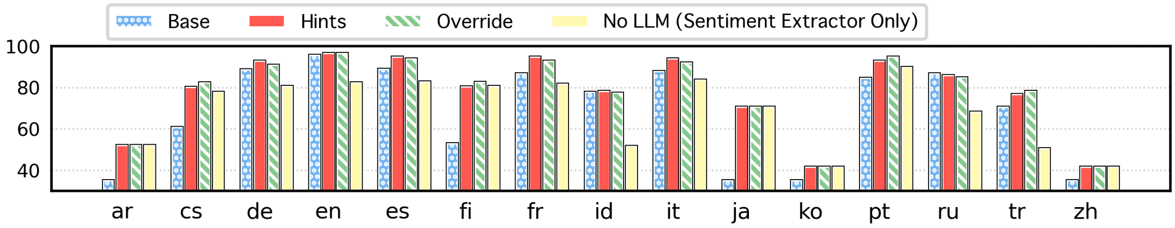

Figure 2: Macro F1 scores of each language on T5, shown with the sentiment extractor's scores in 'No LLM'.

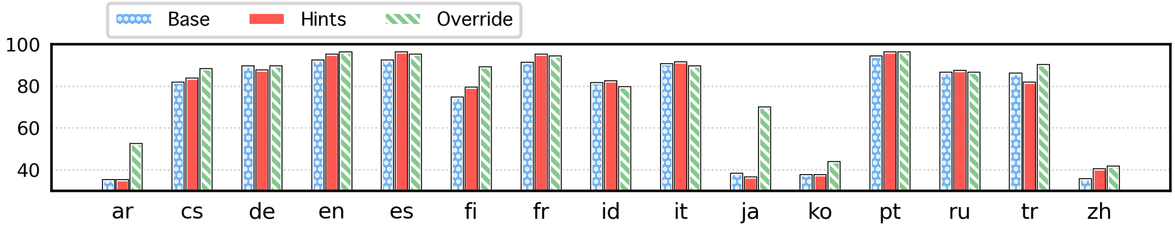

Figure 3: Scores of UL2 model.

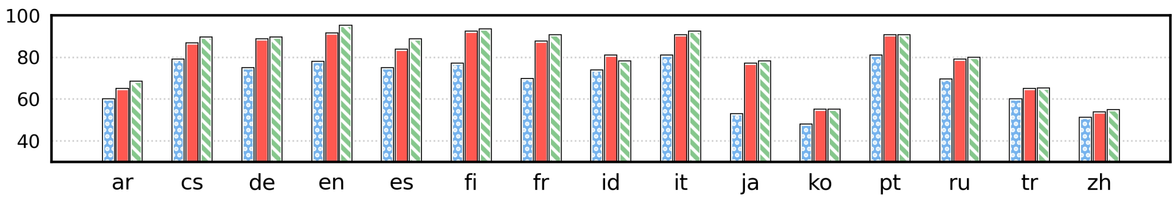

Figure 4: Scores of MT0 model.

guages, but for Arabic, Japanese, Korean and Chinese, the scores are close to the baseline, because the model almost always answers NEGATIVE due to the model's ignorance of words, even characters. The 'Hints' bars show the improvements in 13 languages by incorporating the output of the sentiment extractor, and it was drastic in Czech, Finnish, and Japanese. This was achieved even by the sentiment extractor that is far from perfect for classification, as shown in "No LLM" bars in Figure 2 and their average, 69.4 in Table 1.

Figure 3 shows a similar trend in the UL2 model, but the scores of 'Hints' tend to be lower than 'Override'. This suggests that the model does not recognize the hints added in the method described in Section 3.2 as effectively.

The MT0 model, trained on multilingual data, performs stably for many languages as in Figure 4. Our method was the most effective for MT0 among the six LLMs, especially in French and Japanese, and 33% of errors were reduced in average.

Figures 5 to 7 show that the larger models, Llama2, Mixtral and Llama3 exhibit high multilingual capability in this task, with 'Base' scores of 87 or higher in all languages except for Arabic with Llama2. Even from these high baseline scores, our proposed method improved the scores in 9 languages on Llama2, 12 languages on Mixtral and 7 languages on Llama3. Especially, French with Llama2 and Finnish with Llama3 were improved to near-perfect. Unlike the smaller models, 'Hints' outperformed 'Override' with few exceptions such as Russian with Llama2 and Finnish with Mixtral. This suggests that when the LLM has sufficient knowledge of the language, the best approach is to integrate the LLM and external knowledge, since there is still missing information in the model, and the hints from the external component are appropriately recognized through the prompt.

Indonesian is an outlier that tends to reduce scores from the baseline. It is because the original corpus UD_Indonesian-PUD has significant errors in white-spacing that cause failure of word detection, thus it is difficult for syntax-based component to detect sentiment expressions.

### 5.3 Error Analysis

Errors occurring only in 'Override' showcase the LLM's capability to appropriately ignore hints from external knowledge. Here is an example in German: in "... so gut wie unmöglich geworden"

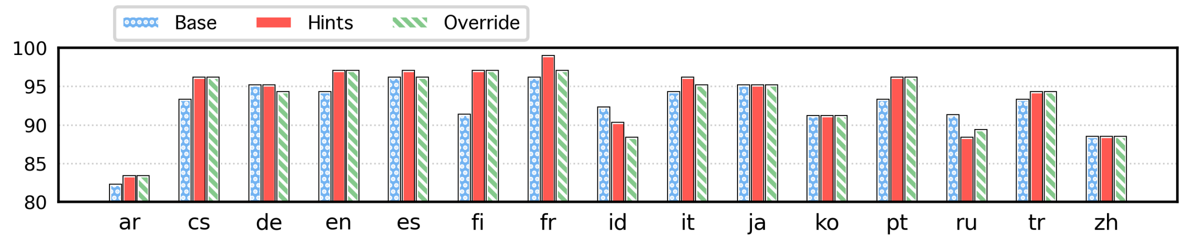


Figure 5: Scores of Llama2 model. Note that Figures 5 to 7 are shown in a different scale from Figures 2 to 4.

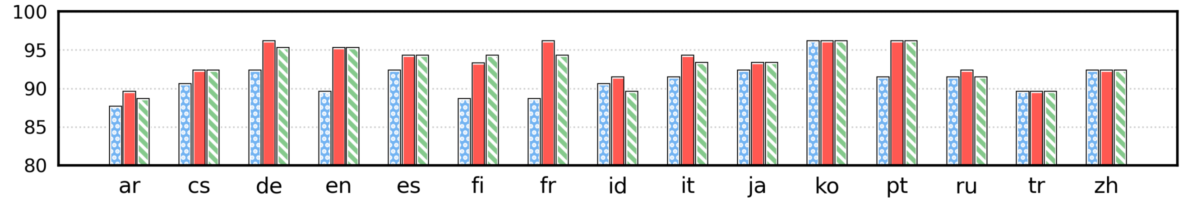


Figure 6: Scores of Mixtral model.

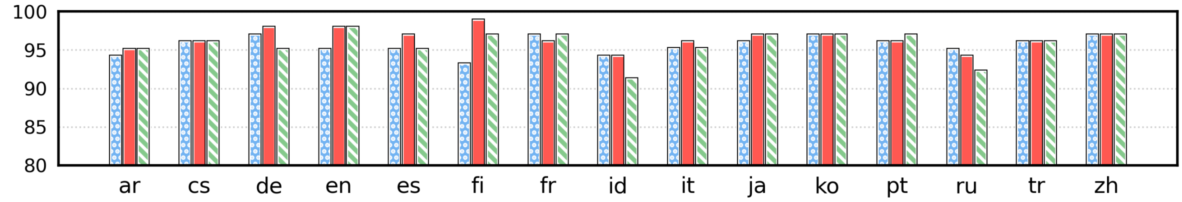


Figure 7: Scores of Llama3 model.

('become almost impossible'), the sentiment extractor detected "gut" ('good') as a positive word but it is not true in this context. MT0 was influenced by the hint and wrongly judged as `POSITIVE`, but other models recognized this idiom and they answered `NEGATIVE` even when an incorrect hint was provided.

On the other hand, errors occurring only in 'Hints' reveal a need for improvement in the hints provided in prompts. MT0 tends to be confused with negation. For example, even with the hint 'hint: "no claro" is a Spanish phrase that has a negative meaning.', the model answered `POSITIVE`. This suggests the model likely misunderstood that "claro" ('clear') was a negative word and interpreted it as negated.

## 6 Conclusion

This paper exploited external knowledge for the multilingual sentiment classification task on LLMs and achieved promising results. It demonstrated that the sentiment extractor can compensate the lack of linguistic knowledge in LLMs in several languages. Additionally, even when incorrect information was provided, the LLMs were able to handle apparent mistakes, suggesting that the combination of the two approaches works best. However, several models did not recognize all the hints, indicating further enhancement of this combination is possible with better prompting techniques.

## 7 Limitations

This paper aims to explore the capability to incorporate external lexical and syntactic knowledge even when an LLM does not perform adequately for certain languages and domains, rather than solely pursuing state-of-the-art scores in a specific classification task. Therefore, there may be better systems (*e.g.* GPT-4) to solve this task more accurately. Nonetheless, this work is meaningful as it aims to improve the task even when large models are not applicable due to computational or legal reasons. Additionally, there is another advantage for users to control the resources without needing to retrain or fine-tune LLMs.

The sentiment extractor utilized in this paper does not always extract correct sentiment expressions and indeed, the recall is low for languages such as Korean and Chinese. Such an external component cannot be perfect but it is possible to enhance the syntactic rules and lexical resources when a new language is needed, and the updated components can naturally be applied to our technique.

Since we are using small parallel data sets and not providing any machine learning model, the prompts and parameters were not fully optimized for this task and data set, except for the additional line for the prompt for the MT0 model described in Section 4.2. Thus there can be better-performing prompts and parameters that could potentially improve the results further. Also, the evaluation datasets with limited number of instances cannot support the results with statistical significance for each language, thus we provide qualitative discussion with interpretable instances with outputs by the sentiment extractor.

## 8 Ethical Statement

Since we are using an existing dataset and LLMs for evaluation and we don't release new text data, there is no ethical concerns in this paper.

## References

AI@Meta. 2024. Llama 3 model card.

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.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Hiroshi Kanayama and Ran Iwamoto. 2020. How Universal are Universal Dependencies? Exploiting Syntax for Multilingual Clause-level Sentiment Detection. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, pages 4063–4073.

Zhoubo Li, Ningyu Zhang, Yunzhi Yao, Mengru Wang, Xi Chen, and Huajun Chen. 2023. Unveiling the pitfalls of knowledge editing for large language models. *arXiv preprint arXiv:2310.02129*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Peng Qi, Timothy Dozat, Yuhao Zhang, and Christopher D. Manning. 2018. Universal dependency parsing from scratch. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 160–170, Brussels, Belgium.

Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. 2023. Text classification via large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8990–9005, Singapore.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations (ICLR)*.

Genta Winata, Lingjue Xie, Karthik Radhakrishnan, Shijie Wu, Xisen Jin, Pengxiang Cheng, Mayank Kulkarni, and Daniel Preotiuc-Pietro. 2023. Overcoming catastrophic forgetting in massively multilingual continual learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 768–777, Toronto, Canada. Association for Computational Linguistics.

Daniel Zeman and et al. 2017. CoNLL 2017 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada.

Yang Zhao, Tetsuya Nasukawa, Masayasu Muraoka, and Bishwaranjan Bhattacharjee. 2023. A simple yet strong domain-agnostic de-bias method for zero-shot sentiment classification. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3923–3931.

ar: سمح الميثاق بإنشاء نقابةٍ للتجار يديرها مواطنو البلدة لفرض ضريبة على من يمر من منطقتهم.
cs: Zakládací listina umožnila vytvoření kupeckého spolku ovládaného městskými zastupiteli, který mohl vybírat daň od těch, co tímto samosprávným městem projížděli.
de: Die Gründungsurkunde ermöglichte die Schaffung einer Händlerinnung, die von den Bürgern der Stadt geführt wurde, um Steuern von durch die Gemeinde reisenden Menschen zu erheben.
en: The charter allowed for the creation of a merchants' guild, run by the town's burgesses to tax people passing through the borough.
es: La carta permitía la creación del gremio de los mercaderes y se gestionaba por los burgueses de la ciudad para cobrar impuestos a las personas a través del municipio.
fi: Peruskirjassa sallittiin kaupungin porvareiden johtaman kauppiaiden killan luominen, jonka tarkoituksena oli verottaa ihmisiä, jotka kulkivat kauppalan läpi.
fr: La charte permit la création d'une guilde de marchands, dirigée par la bourgeoisie de la ville, qui faisait payer un impôt aux personnes traversant le quartier.
id: Piagam ini mengizinkanpembentukanserikat pedagang, yang dijalankanoleh perwakilankota untuk menarik pajak dari orang yang melintasiborough.
it: Il trattato permise la creazione di una corporazione dei mercanti, gestita dai deputati della città per tassare coloro che attraversavano la zona,
ja: この憲章は、都市の選出代議士が商人向けの組合を設立することを認めました。
ko: 상인 협회의 수립이 선언문을 통해 가능해졌는데협회는 자치구를 통과하는사람들을 대상으로 세금을 징수하는마을의 대의원들에 의해 운영되었다.
pt: A carta permitiu a criação de uma guilda de comerciantes, administrada pelos burgueses da vila, para tributar as pessoas que passavam pelo bairro.
ru: Этот устав позволил создать организацию торговцев, руководимую жителями города для обложения налогом людей, проходящих через городок.
tr: Tüzük, kasabadan geçen kişilerin vergilendirilmesi için kasaba sakinleri tarafından yönetilen bir tüccar locası oluşturulmasını sağladı.
zh: 憲章曾允許創立由城鎮市民管理的商人行會，以向途經城鎮的人士徵收稅費。

Figure 8: Example from Parallel Sentiment and extraction results by the sentiment extractor. POSITIVE label was commonly given for the parallel sentences in 15 languages. Highlighted words in blue are the positive words detected by the sentient extractor, and the underlined words are the targets of the positive sentiment (targets are not cared in prompting in this work).

| | | T5 | | UL2 | | MT0 | | Llama2 | | Mixtral | | Llama3 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | hints | − | + | − | + | − | + | − | + | − | + | − | + |
| ar | | N | N | N | N | P | P | N | N | P | P | P | P |
| cs | $p$ | N | P | P | P | P | P | N | P | P | P | P | P |
| de | $p$ | N | P | P | P | P | P | P | P | P | P | P | P |
| en | | P | P | P | P | P | P | N | N | P | P | P | P |
| es | $p$ | N | P | N | P | N | P | N | P | P | P | P | P |
| fi | $p$ | N | N | N | N | N | P | N | P | P | P | N | P |
| fr | $p$ | N | P | P | P | N | P | N | P | P | P | P | P |
| id | | N | N | N | N | P | P | N | N | P | P | P | P |
| it | $p$ | N | P | P | P | P | P | N | P | P | P | P | P |
| ja | | N | N | N | N | P | P | P | P | P | P | P | P |
| ko | | N | N | N | N | P | P | N | N | P | P | N | N |
| pt | $p$ | N | P | N | P | P | P | N | P | P | P | P | P |
| ru | | N | N | N | N | P | P | N | N | N | N | P | P |
| tr | | N | N | P | N | P | P | P | P | P | P | P | P |
| zh | | N | N | N | N | P | P | N | N | P | P | P | P |

Table 2: Hints generated by the sentiment extractor, and sentiment classification results without and with hints for the sentences in Figure 8 on the four models. Highlighted cells indicate the correct predictions.

# A Appendix

## A.1 Prompt for Phrase

The hint prompt shown in Section 3.2 is for a sentiment expression consisting of a single word with a polarity. When multiple words form a positive or negative expression, the hint is generated in the following format:

> hint: "[phrase]" is a(n) [lang] phrase which has a [polarity] meaning.

For example, a sentence 'we never like it' has a negative expression, though "like" is a positive verb[7]. The hint in this case is:

> hint: "never like" is an English phrase which has a negative meaning.

Similarly "have merit" (positive) and "too short" (negative) are given as phrasal hints.

## A.2 Prompt Adjustment

As mentioned in Section 4.2, the MT0 model tends to output POSITIVE. Therefore, we replace "Please just answer the label." in the prompt in Section 3.1 with the following sentence. This increases the MT0's base scores by 20% on average.

> In this case all input is either of them, and do not hesitate to select NEGATIVE when you think the input is relatively negative things in some way.

## A.3 Multilingual Examples

Figures 8 and 9 show the examples in Parallel Sentiment. As shown in the result of automatic sentiment extraction, the extractor does not always

[7]The sentiment annotator cares about the parts-of-speech, thus "like" used as a preposition is not regarded as a positive word.

ar: نقل أن تيبيريوس أبدى ندمه على رحيله وطلب أن يعود إلى روما عدة مرات, لكن في كل مرة كان أغسطس يرفض طلبه.
cs: Tiberius podle dostupných pramenů svého odjezdu litoval a několikrát žádal, aby se směl do Říma vrátit, ale Augustus jeho žádosti pokaždé zamítl.
de: Berichten zufolge bereute Tiberius seine Abreise und bat mehrfach um Erlaubnis, nach Rom zurückzukehren, doch Augustus verweigerte ihm jedes Mal seine Bitte.
en: Tiberius reportedly regretted his departure and requested to return to Rome several times, but each time Augustus refused his requests.
es: Supuestamente, Tiberio se arrepintió de haberse marchado y solicitó varias veces regresar a Roma, pero Augusto rechazó todas sus solicitudes.
fi: Tiberiuksen kerrotaan katuneen lähtöään ja pyytäneen monta kertaa saada palata Roomaan, mutta joka kerran Augustus hylkäsi hänen pyyntönsä.
fr: Tibère aurait regretté son départ et demandé à retourner à Rome plusieurs fois, mais Auguste aurait refusé chacune de ses demandes.
id: Tiberius dilaporkanmenyesalikeberangkatannya dan meminta untuk kembali ke Roma beberapa kali, namun Augustus menolak permintaannya setiap kali.
it: Presumibilmente, Tiberio si pentì della sua partenza e richiese di tornare a Roma molte volte, ma Augusto rifiutò le sue richieste ogni volta.
ja: ティベリウスは自身がローマを離れたことを後悔し、ローマに戻ることを何度か要求したと伝えられているが、その都度、アウグストゥスがその要求を拒否していた。
ko: 티베리우스는 자신이 떠난 것에 대해 후회하고여러 차례 로마로 불러달라고 요청했으나그 때마다 아우구스투스는 티베리우스의 요청을 거절하였다.
pt: Tibério supostamente lamentou a sua partida e pediu para regressar a Roma várias vezes, mas Augusto sempre recusou seus pedidos.
ru: Тиберий, по имеющимся сведениям, пожалел о своем уходе и просил несколько раз вернуться в Рим, но Август каждый раз отказывал его просьбам.
tr: Aktarılana göre Tiberius, gidişinden pişman oldu ve birkaç kez Roma'ya dönmek istedi ancak Augustus taleplerini her seferinde reddetti.
zh: 據稱提庇留離開後感到後悔，並數次要求返回羅馬，但是均被奥古斯都拒絕。

Figure 9: Another example with NEGATIVE annotation. The words highlighted in pink are negative words detected by the system. In Spanish, a preterite form of "solicitar" ('to request') was wrongly detected as positive due to a wrong parsing result in a coordination structure.

| | hints | T5 | | UL2 | | MT0 | | Llama2 | | Mixtral | | Llama3 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | − | + | − | + | − | + | − | + | − | + | − | + |
| ar | | N | N | N | N | P | P | N | N | N | N | N | N |
| cs | | N | N | N | N | N | N | N | N | N | N | N | N |
| de | | N | N | N | P | P | P | N | N | N | N | N | N |
| en | $n$ | N | N | N | N | P | N | N | N | N | N | N | N |
| es | $n$,$p$ | N | N | N | N | P | P | N | N | N | N | N | N |
| fi | | N | N | N | N | N | N | N | N | N | N | N | N |
| fr | $n$ | N | N | N | N | P | N | N | N | N | N | N | N |
| id | $n$ | N | N | N | N | P | N | N | N | N | N | N | N |
| it | $n$ | N | N | N | N | P | N | N | N | N | N | N | N |
| ja | $n$, $n$ | N | N | N | N | P | N | N | N | N | N | N | N |
| ko | | N | N | N | N | P | P | N | N | N | N | N | N |
| pt | $n$,$n$ | N | N | N | N | N | N | N | N | N | N | N | N |
| ru | $n$ | N | N | N | N | P | N | N | N | N | N | N | N |
| tr | | N | N | N | N | P | P | N | N | N | N | N | N |
| zh | | N | N | N | N | P | P | N | N | N | N | N | N |

Table 3: Results for sentences in Figure 9. Highlighted cells indicate the correct predictions.

extract sentiment expressions, and the recall is low in some languages such as Arabic, Korean and Chinese. It causes the low scores shown in "No LLM" bars in Figure 2.

Table 2 shows the sentiment classification results for the sentences in Figure 8. The column labeled 'hints' shows positive ($p$) or negative ($n$) words provided as hints, which correspond to highlighted words in Figure 8. The right part shows the classification results by six models, without ('$-$') or with ('$+$') hints. 'P' and 'N' denote POSITIVE and NEGATIVE respectively. 'P' is the correct label for all the languages, so we can see the classification by all models was changed to the correct one in one or more languages, and when no hint was given, the result was unchanged, with an exception due to nondeterministic behavior of the UL2 model.

Table 3 shows the results for another example in Figure 9. The correct label is 'N'. MT0 tends to select POSITIVE but the hints successfully invert the decision in 6 languages. The sentiment extractor outputs both positive and negative expressions in Spanish and both were added as hints, but it did not have detrimental results. This was the easy case for other models so prediction was NEGATIVE even without hints.