

Findings of the WASSA 2024 EXALT Shared Task on Explainability for Cross-Lingual Emotion in Tweets

Aaron Maladry, Pranaydeep Singh and Els Lefever

LT3, Ghent University, Belgium

{aaron.maladry,pranaydeep.singh,els.lefever}@ugent.be

Abstract

This paper presents a detailed description and results of the first shared task on explainability for cross-lingual emotion in tweets. Given a tweet in one of the five target languages (Dutch, Russian, Spanish, English, and French), systems should predict the correct emotion label (Task 1), as well as the words triggering the predicted emotion label (Task 2). The tweets were collected based on a list of stop words to prevent topical or emotional bias and were subsequently manually annotated. For both tasks, only a training corpus for English was provided, obliging participating systems to design cross-lingual approaches. Our shared task received submissions from 14 teams for the emotion detection task and from 6 teams for the trigger word detection task. The highest macro F1-scores obtained for both tasks are respectively 0.629 and 0.616, demonstrating that cross-lingual emotion detection is still a challenging task.

Emotion detection is a well-studied task in the field of NLP and has already been addressed in previous SemEval shared tasks (Mohammad et al., 2018; Chatterjee et al., 2019). In this shared task, however, we wanted to go one step further and offer a manually annotated multilingual benchmark data set, where not only emotions are labeled, but also the words triggering these emotions. To this end, we aim to investigate to what extent emotion information is transferable across languages, by offering training data in English, and evaluation data for 5 different target languages, namely Dutch, Russian, Spanish, English, and French. In addition, predicting trigger words should be a first step to endorsing emotion detection systems with a means to explain why a specific emotion has been predicted. With an ever-rising flurry of black-box models, we aim to foster research that moves towards the interpretability and explainability of systems.

As there is no real consensus on a standard emo-

tion labeling framework, we opted to apply the label set of Debruyne et al. (2019), which is justified both theoretically and practically. Frequency and cluster analysis of tweet annotations resulted in a label set containing 5 emotions: *Love*, *Joy*, *Anger*, *Fear* and *Sadness*. This label set certainly shows a resemblance to Ekman’s basic emotions, but due to the applied data-driven approach, the label set is more grounded in the task of emotion detection in social-media text. As we did not crawl our data based on emojis (as opposed to Debruyne et al.), we also added a *neutral* emotion label to the emotion label set.

1 Dataset Construction

We present a fully annotated dataset of multilingual tweets that were gathered using the Twitter API. The dataset includes a variety of 5 European languages, including Dutch, Russian, Spanish, English and French. For each of these languages, we collected the tweets based on a list of stop words that originate from either Spacy (Honnibal and Montani, 2017) or NLTK (Bird et al., 2009) depending on the availability for the language. The stopwords were subsequently manually filtered by native speakers to remove any incorrect entries or content words. The selected tweets do not target a specific time frame, although we did exclude the COVID-19 years 2019-2021 to avoid a topical bias. With the use of stop words, the collected data is less likely to be affected by specific topical or emotional biases that can be inherited from search terms. After collecting a significant batch of about 200,000 tweets (excluding retweets), we randomly selected subsets to annotate for each language.

For the emotion labeling process, the annotators were provided with detailed annotation guidelines (Singh et al., 2023). The first level of annotations describes only the primary emotion expressed by the text, for which we assume the perspective

of the reader and answer the question “Which emotion do you think the writer intends to convey?”. This means we only employ a single emotion label. As mentioned before, the emotion labels we use here are Love, Joy, Fear, Sadness, Anger and Neutral. These labels were the outcome of a broad study (De Bruyne et al., 2019), where annotations for 25 specific emotion labels were clustered into 5 broader categories, to which we added the neutral class. In the annotation guidelines, we provided all labels that are grouped under this broader emotion class as helper labels. The label “anger”, for instance, groups the fine-grained emotion labels “disgust”, “frustration”, “rage” and “resentment”. Inspecting these fine-grained labels can help the annotator to confidently choose the coarse-grained emotion label (“anger”). All annotators indicated they found this helpful when deciding between two similar positive emotions, such as “love” and “joy”. As the collected tweets are gathered in a manner that aims to collect as generic data as possible, this also resulted in a lot of noise. Many tweets rely on specific contextual information, and as a result, some tweets cannot be interpreted and annotated reliably. Moreover, many tweets are obvious “spam” tweets, posted by automated bots. For these tweets, we introduced a “discard” label as an addition to the emotion label set. The tweets labeled with this discard label, which is around 30% of the annotated data, are excluded from the shared task to guard the quality of the annotations. Figure 1 illustrates the annotation of the emotion labels per tweet.

The screenshot shows a tweet with the text "Stay away from me and mines or u gonna get hurt." Below the tweet is a form titled "Emotion label". The form contains several radio button options: Anger^[1] (checked), Sadness^[2], Fear^[3], Joy^[4], Love^[5], Neutral^[6], and Discard^[7].

Figure 1: Sub-task 1: Annotation of Emotion label

To gauge the inter-annotator agreement for our labeling scheme, we tested the annotation scheme for Dutch, which is the native language of our 5 annotators. On a subset of 50 samples, we calculated Fleiss’ Kappa (Fleiss, 1971), resulting in a moderate agreement score of 0.62, which can be considered satisfactory given the subjective nature of this task. Moreover, this agreement study includes the tweets that were annotated with the “discard”

label. For the shared task train and test data, we remove these discards and therefore expect higher agreement on the actual emotion labels.

In addition to these primary emotion labels, the annotators were also instructed to identify the words that evoke that emotion. As these are the words that “trigger” the emotion, we refer to them as **trigger words** (Schroth et al., 2005). As a guideline to aid in identifying these trigger words, we instructed the annotators to imagine the text without one or more of the words. If the emotion changes or disappears when the words are removed or replaced, it points to the vitality of these words in identifying the emotion. Emojis, emoticons, and punctuation (such as ... and ! or ???) can also be indicated as a trigger. Detailed information on how these trigger words have been labeled can be found in the annotation guidelines (Singh et al., 2023). Figure 2 shows an example of the trigger words selected for a specific tweet.

The screenshot shows a tweet with the text "Stay away from me and mines or u gonna get hurt." Above the tweet is a form titled "Trigger words". The form contains a text input field with the value "Trigger 0" and a button labeled "Stay away" which is highlighted in blue, indicating it has been selected as a trigger word.

Figure 2: Sub-task 2: Annotation of trigger words.

We conceptualize trigger word detection as a binary token classification task, and calculate evaluation metrics in a pairwise manner, taking each annotator as the gold standard. For agreement on span detection tasks, Mathet et al. (2015) propose a gamma metric for unified class- and span agreement. More specifically, we employed soft gamma, which allows multiple-span annotations of an annotator to be matched with a single continuous span of another annotator. This makes sense for this task because it does not matter whether the important words are annotated as a continuous span or split into multiple shorter spans as long as they still correspond to the larger span. Using this metric, we calculated the agreement for each sentence and then averaged it to attain a corpus-level agreement score. This results in a soft gamma score of 0.4383, indicating that this task is quite subjective but still shows fair agreement (with a 0 gamma score representing random agreement and a 1 indicating complete agreement). Whilst this metric is created for span agreement in particular, there

are a few arguments against its use for our trigger word detection task. Firstly, averaging across all sentences somewhat defeats the purpose of using a chance-corrected agreement metric because the agreement is only calculated on a sentence level and not on a corpus level. Secondly, it would be possible to calculate this agreement on the corpus level by concatenating the sentences, but in that case, longer sentences would have a higher impact on the score than shorter sentences, which is not a desired effect. To circumvent these limitations, we decided to evaluate the agreement between annotators based on Hamming distance and Mean Average Precision. *Hamming distance* is an edit distance metric that counts the words with wrongly assigned labels relative to the number of tokens in a sentence. In addition, we use *Mean Average Precision*, a popular metric for the evaluation of feature importance attribution techniques (Atanasova et al., 2020). Table 1 shows the agreement of trigger word annotations on the Dutch and English sets. The scores for Hamming distance are very small, which indicates that the span annotations are overall rather similar and that no annotators use significantly more trigger words than others. At the same time, the modest MAP scores indicate that the exact words that are essential for one annotator, may not be as essential for the others.

	Hamming	MAP
Dutch Pairwise	0.04	0.13
English Pairwise	0.03	0.18

Table 1: Trigger word annotation agreement

2 Tasks and Evaluation

Based on the manually annotated data set, we present the following two sub-tasks.

2.1 Cross-lingual Emotion Detection Task

The first task is to predict the correct emotion label for each tweet from 6 possible classes: *Love, Joy, Anger, Fear, Sadness, Neutral* in five target languages. More concretely, we present the participants with a train set of 5000 English tweets, a multi-lingual development set of 500 tweets and a multi-lingual test set of 2500.¹ The participants are free to use additional training resources, though

¹All data is publicly available after registration through <https://huggingface.co/datasets/pranaydeeps/EXALT-v1>.

they should be restricted to English only to evaluate the efficacy of the cross-lingual setup. Evaluation of this classification task is performed through macro-averaged precision, recall, and F1-score metrics, with the systems ranked based on their F1-score.

2.2 Prediction of the text span triggering the predicted emotion label

For the second task, we propose trigger word detection, a task that is focused on explaining which words are used to express the emotion. For the trigger word detection task, we discarded tweets that did not contain any trigger words, which leaves us with a train set of 3000 English tweets, a multi-lingual development set of 300 samples, and a multi-lingual test set of 832 samples. To evaluate the scores of the systems of our participants, we considered two evaluation methods. Firstly, we can evaluate trigger word detection as a binary token classification task. We propose using macro-averaged token F1-score for this purpose to overcome the label imbalance (with most tokens not being trigger words). In addition, as we anticipate the use of post-hoc explainability algorithms, we expect some of the authors to generate numerical importance tokens instead of binary indicators for each word. To evaluate these numerical importances, we employ Accumulated Precise Importance attribution, or API-score (Maladry et al., 2024). This is a sentence-level metric that sums up the (normalized) numerical importance of each trigger word token (based on human annotations). To illustrate, the API score for Example 1 equals 95% (33 + 18 + 19 + 25), with only 5% of the total importance being attributed to a non-trigger word.

Example 1

	<i>sad about my rejected paper :(</i>					
HUM.	1	0	0	1	1	1
IMP.	.33	0	.05	.18	.19	.25

3 Applied Methodologies

3.1 Baseline

Our baseline model for cross-lingual emotion detection employs an XLM-RoBERTa model fine-tuned on the English training data. As the ideal system setup makes use of a single model and can provide ad-hoc explainable predicts, we continued from the same fine-tuned emotion detection model for trigger word detection and applied the Layer

Integrated Gradients (Sundararajan et al., 2017) algorithm with transformers-interpret² to generate sub-token importance.³ After matching these sub-token importance attributions back to the original words and normalizing them, we achieved the predictions for our numerical trigger word detection task. We converted all word-level importances to binary vectors based on the numerical values. All words with an importance value above 10% (compared to the total importance attributions in the sentence) were considered trigger words based on an exploratory study on a validation set. This conversion is illustrated in Example 2.⁴

Example 2

	<i>sad about my rejected paper :(</i>					
HUM.	1	0	0	1	1	1
IMP.	.33	0	.05	.18	.19	.25
$x=10\%$	1	0	0	1	1	1

3.2 Participating Teams

As shown in Table 2, most participants to our shared task used fine-tuned large generative models like GPT4 and GPT3.5 (OpenAI et al., 2024), Gemma (Team et al., 2024), LLaMa-3 (AI@Meta, 2024), etc.

Besides the data we provided for the shared task, many participants also used external resources, such as earlier shared tasks for emotion detection (Mohammad et al., 2018; Chatterjee et al., 2019), and data augmentation methods. These augmentation methods often include translation to the target languages. Backtranslation was also a frequently employed method to augment the data.

For the trigger word detection tasks most teams started with a fine-tuned token classification system to calculate the numerical scores for the numerical trigger word detection task. In contrast, some systems started from a classification model and employed the same importance attributions techniques used by the baseline model. Although many parameters, thresholds, and different feature importance algorithms can be experimented with, the participants did not explore this extensively.

Some of the more distinct approaches for the shared task employed multi-agent work-

²<https://github.com/cdpierse/transformers-interpret>

³The code (and models) used for the baselines is publicly available through https://github.com/pranaydeeps/WASSA24_EXALT/tree/main/starters_kit.

⁴This example originates from Maladry et al. (2024), where x signifies the chosen threshold for converting a numerical attribution to a binary indicator.

flows (Cheng et al., 2024a) and label projection with trigger word switching (Šmíd et al., 2024).

4 Results

4.1 Emotion Detection

As shown in Table 3, all 14 submitting teams outperformed the baseline score of 0.4476 macro-F1 for the emotion detection sub-task. Team 1024m (Kadiyala, 2024) had the best-performing system by quite a big margin with a macro-F1 of 0.6295 on the test set. The team utilized an ensemble of Gemma (Team et al., 2024), Claude-Opus⁵, Llama-3 (AI@Meta, 2024), Mistral-v2-7B (Jiang et al., 2023) and applied majority voting for the final prediction. The second-best system, Team BCSZ (Cheng et al., 2024a), also employed an ensemble of generative models with the addition of different varieties of Agentic Workflows where an additional decision-making LM is deployed to make the final prediction based on the output of the individual LMs. Team DKE-Research (Wang et al., 2024) submitted the best-performing system without using large generative LMs directly or indirectly, with a macro-F1 of 0.5661. The team utilized knowledge distillation by training a monolingual English teacher model for emotion and transferring the knowledge to a multilingual model.

4.2 Binary Trigger Word detection

For the second sub-task of Binary Trigger Word Detection, again all 6 submitting teams comprehensively beat the baseline of sub-token importance using Layer Integrated Gradients. Team CT-cloud (Zhang et al., 2024) had the best-performing system (0.6158) by a small margin of 0.0063 macro-F1. The team applied token classification at the sub-word level, using the highest confidence among the sub-words as the confidence of each word. Most teams used similar post-hoc explainability approaches, while some utilized multi-task learning techniques. Team NYCU-NLP (Lin et al., 2024) used an ensemble of two large generative LMs (Starling-7B (2023) and Llama3-8B (2024)) with instruction fine-tuning.

4.3 Numerical Trigger Word Detection

The leader board of the Numerical Trigger Word Detection sub-task was identical to the Binary Triggers sub-task with one exception. Team UWB (Šmíd et al., 2024), which had the 3rd best

⁵<https://www.anthropic.com/claude>

Team Name	Emotion Rank	Binary Triggers Rank	Numerical Triggers Rank	Large Generative Language Models	Fine-tuning	Translation	Augmentation	Joint Modelling
1024m	1	-	-	✓	✓	✗	✗	✗
BCSZ	2	6	6	✓	✓	✗	✗	✗
Treehouse	3	-	-	✓	✗	✗	✓	✗
NYCU-NLP	4	5	5	✓	✓	✗	✗	✓
HITSZ-HLT	6	2	3	✗	✓	✓	✗	✓
UWB	7	3	1	✓	✓	✓	✗	✗
wu_tlaxe	8	-	-	✓	✓	✓	✗	✗
DKE-Research	9	-	-	✗	✓	✓	✗	✗
NLPNewcomer	10	4	4	✗	✓	✗	✗	✓
CTcloud	11	1	2	✗	✓	✗	✓	✗
PCICUNAM	12	-	-	✗	✓	✓	✗	✗
LLiiMas	13	-	-	✗	✓	✓	✗	✗
EXALT-Baseline	15	7	7	✗	✓	✗	✗	✗

Table 2: An overview of the methodologies used by the teams for the shared task and their overall rankings on the respective leaderboards of each sub-task.

Team	Emotion_F1	Team	Token F1
1024m (Kadiyala, 2024)	0.6295	CTcloud (Zhang et al., 2024)	0.6158
BCSZ (Cheng et al., 2024a)	0.6046	HITSZ-HLT (Xiong et al., 2024)	0.6095
Treehouse (Cheng et al., 2024b)	0.6015	UWB (Šmíd et al., 2024)	0.5919
NYCU-NLP (Lin et al., 2024)	0.5951	NLPNewcomer	0.5785
CTYUN-AI	0.5911	NYCU-NLP (Lin et al., 2024)	0.5636
HITSZ-HLT (Xiong et al., 2024)	0.591	BCSZ (Cheng et al., 2024a)	0.4778
UWB (Šmíd et al., 2024)	0.591	EXALT-Baseline	0.2349
wu_tlaxe (Davenport et al., 2024)	0.573		
DKE-Research (Wang et al., 2024)	0.5661		
NLPNewcomer	0.5444		
CTcloud (Zhang et al., 2024)	0.5428		
PCICUNAM (Vázquez-Osorio et al., 2024)	0.5183		
LLiiMas	0.5067		
(Vázquez-Osorio and Gómez-Adorno, 2024)			
EXALT-Baseline	0.4476		

Table 3: Leader board based on macro-averaged F1-scores for Emotion Detection

system for the binary task, achieved 1st place for the numerical triggers based on the API (Accumulated Precise Importance) metric. The methodology involved translating the English data into the target languages to generate additional training data while using special symbols for the trigger words to transfer them to the target language. They also utilize trigger-word switching, i.e., swapping trigger words between an English sentence and a translated sentence in one of the target languages. These simple yet ingenious ideas led to the best-performing system with an API-score of 0.7052.

5 Discussion

For this shared task, all teams experimented with widely varying methods for system fine-tuning and prompting large generative LMs (see Table 2), and these approaches have also resulted in some of the best systems for the first sub-task. The top 8 teams have directly or indirectly (Team HITSZ-HLT has indirectly used ChatGPT for augmenting their data) employed generative LMs. It is, how-

Table 4: Leader board based on macro-averaged token F1-score for trigger word detection.

Team	API-score
UWB (Šmíd et al., 2024)	0.7052
CTcloud (Zhang et al., 2024)	0.6972
HITSZ-HLT (Xiong et al., 2024)	0.6961
NLP_Newcomer	0.658
NYCU-NLP (Lin et al., 2024)	0.6442
BCSZ (Cheng et al., 2024a)	0.4548
EXALT-Baseline	0.216

Table 5: Leader board based on Accumulated Precise Importance for trigger word detection.

ever, surprising to see limited experimentation with approaches tailored for cross-lingual tasks such as MAD-X (Pfeiffer et al., 2020) or BLOOMZ (Muenighoff et al., 2023).

Figures 3 and 4 also visualize some interesting findings for individual emotion labels in the test set, as well as for each target language. *Neutral* and *Anger* seem to be the easiest emotions to label, while *Fear* is often the hardest. This can be attributed to the class imbalance as *Fear* had the least samples in the train and test sets by a significant margin. For the languages, surprisingly, English is not the best-performing target language. All systems performed best on the Spanish test set, while

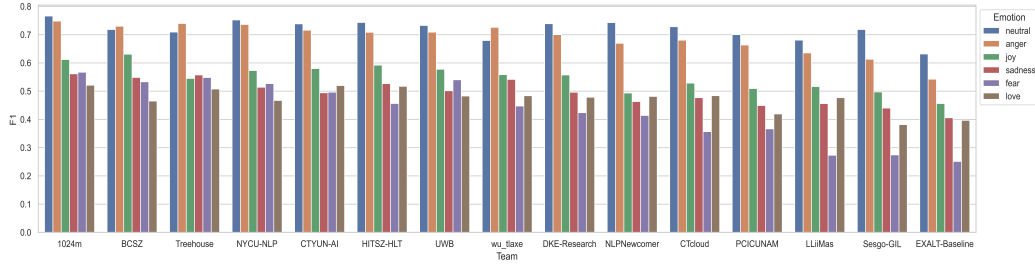


Figure 3: Scores per emotion label

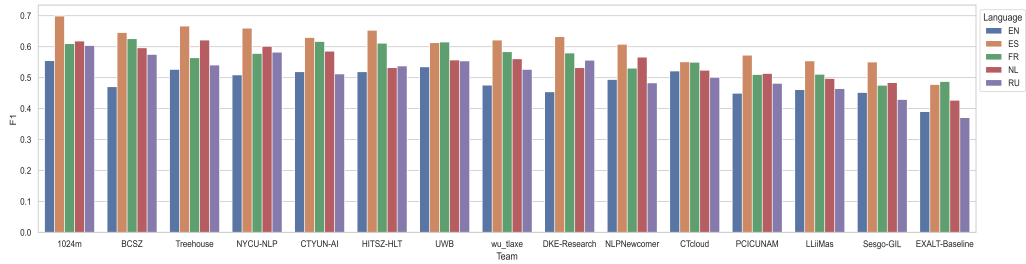


Figure 4: Scores per language

English and Russian were often the hardest. For detailed scores per emotion and per language for each system, please refer to Table 6.

It is also worthwhile to note that none of the teams experimented with alternate methodologies for sub-word attributions which is considered one of the cornerstones of explainability for LLMs or other ad-hoc explainability ideas. One of the more unique approaches for trigger word detection was the trigger word switching for data augmentation, applied by Team UWB, which relies on specifically translating (some or all) trigger words to the target languages. A few teams (NYCU-NLP, HITSZ-HLT, NLPNewcomer) employed joint multi-task approaches for the 3 sub-tasks. While these approaches are somewhere between post-hoc and ad-hoc interpretability, and therefore more interpretable than most of the post-hoc approaches proposed, they did not result in the top-performing system for the trigger-word detection sub-tasks.

6 Conclusion & Future work

In this shared task, we were able to fully work out our annotation scheme and streamline the annotation process for labeling emotion and trigger words for emotion. Thanks to our participants, we could investigate the cross-lingual transfer from English to a variety of European languages. Whilst cross-lingual transfer remains a relevant research topic,

transferring to some languages is easier than others. As shown by related work, the difficulty of language transfer can be affected by different factors, such as the relatedness of the source and target language (de Vries et al., 2022) and whether or not the two languages share the same script (Zubillaga et al., 2024). The quantity of available resources for both languages, but especially the target language is highly relevant. This leaves us with three particular challenges: language relatedness, the availability of a shared script, and the availability of language resources. In the current run of the task, this was kept in check by selecting only European languages that (1) have sufficient resources, (2) are closely related and (3) share the same Latin script. Therefore, one of our goals for future runs of the shared task would be to expand the evaluation pool to languages that are strongly impacted by these difficulties.

In addition, whilst the approaches for cross-lingual explainability for emotion detection work quite well, one of the downsides remains that they are restricted to post-hoc explainability, often employing a separate system that is trained specifically for the explanation but is not explainable intrinsically. For that reason, we suggest future work to investigate the development of inherently (ad-hoc) explainable approaches.

Team	EN	ES	FR	NL	RU	neutral	joy	sadness	anger	fear	love
I024m	0.55	0.70	0.61	0.62	0.60	0.77	0.61	0.56	0.75	0.57	0.52
BCSZ	0.47	0.65	0.63	0.60	0.58	0.72	0.63	0.55	0.73	0.53	0.47
Treehouse	0.53	0.67	0.56	0.62	0.54	0.71	0.55	0.56	0.74	0.55	0.51
NYCU-NLP	0.51	0.66	0.58	0.60	0.58	0.75	0.57	0.51	0.74	0.53	0.47
CTYUN-AI	0.52	0.63	0.62	0.59	0.51	0.74	0.58	0.49	0.72	0.50	0.52
HITSZ-HLT	0.52	0.65	0.61	0.53	0.54	0.74	0.59	0.53	0.71	0.46	0.52
UWB	0.53	0.61	0.62	0.56	0.55	0.73	0.58	0.50	0.71	0.54	0.48
wu_tlaxe	0.48	0.62	0.58	0.56	0.53	0.68	0.56	0.54	0.73	0.45	0.48
DKE-Research	0.45	0.63	0.58	0.53	0.56	0.74	0.56	0.50	0.70	0.42	0.48
NLPNewcomer	0.49	0.61	0.53	0.57	0.48	0.74	0.49	0.46	0.67	0.41	0.48
CTcloud	0.52	0.55	0.55	0.52	0.50	0.73	0.53	0.48	0.68	0.36	0.48
PCICUNAM	0.45	0.57	0.51	0.51	0.48	0.70	0.51	0.45	0.66	0.37	0.42
LLiiMas	0.46	0.55	0.51	0.50	0.46	0.68	0.52	0.46	0.64	0.27	0.48
Sesgo-GIL	0.45	0.55	0.48	0.48	0.43	0.72	0.50	0.44	0.61	0.27	0.38
EXALT-Baseline	0.39	0.48	0.49	0.43	0.37	0.63	0.46	0.41	0.54	0.25	0.40

Table 6: F1-scores for Emotion detection with fine-grained analysis for individual languages and emotion labels.

Limitations

Emotion detection and explaining which part of a text is expressing the emotion are both highly subjective tasks. This is also reflected by the IAA scores included in this study. In our work, we are limited to single-annotator labels and were not able to provide multiple perspectives because that would significantly increase the annotation workload. Furthermore, we would need to find more annotators for all languages, which was not practically feasible. In addition, our annotation scheme does not allow multi-label classification, meaning that texts that contain multiple emotions were either labeled by their primary, most present, emotion or had to be discarded when there was a mix. For future work, we believe it would be interesting to take a closer look at the tweets that were discarded in this study to distinguish multi-emotion tweets from tweets that could not be understood without context. Furthermore, our study is still limited to European languages, which are relatively closely related. We expect that transferring models for emotion detection to unrelated or more distantly related languages is significantly harder and poses an interesting avenue for future research.

Acknowledgments

Our special thanks go out to the student workers who performed the emotion annotation work for this shared task: Cas Van Raemdonck, Febe Coussée, Fran Meulebrouck, Jef Naert, Robbe Verwerft, and Wout Boven. This work was supported by Ghent University under grant BOF.24Y.2021.0019.01 .

References

- AI@Meta. 2024. [Llama 3 model card](#).
- Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. [A diagnostic study of explainability techniques for text classification](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online. Association for Computational Linguistics.
- Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc."
- Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal. 2019. [SemEval-2019 task 3: EmoContext contextual emotion detection in text](#). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 39–48, Minneapolis, Minnesota, USA. Association for Computational Linguistics.
- Long Cheng, Qihao Shao, Christine Zhao, Sheng Bi, and Gina-Anne Levow. 2024a. [Teii: Think, explain, interact and iterate with large language models to solve cross-lingual emotion detection](#). In *Proceedings of the 14th Workshop of on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis@ACL 2024*, Bangkok, Thailand.
- Yao-Fei Cheng, Jeongyeob Hong, Andrew Wang, Anita Silva, and Gina-Anne Levow. 2024b. [Treehouse at wassa 2024 shared task2: Explainability for cross-lingual emotion in tweets](#). In *Proceedings of the 14th Workshop of on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis@ACL 2024*, Bangkok, Thailand.
- Jon Davenport, Keren Ruditsky, Anna Batra, and Yulha Lhawa. 2024. [Wu_tlaxe at wassa 2024 explainability for cross-lingual emotion in tweets shared task 1: Emotion through translation using twhin-bert and gpt](#). In *Proceedings of the 14th Workshop of on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis@ACL 2024*, Bangkok, Thailand.

- Luna De Bruyne, Orphée De Clercq, and Veronique Hoste. 2019. Towards an empirically grounded framework for emotion analysis. In *Proceedings of HUSO 2019, the fifth international conference on human and social analytics*, pages 11–16. IARIA, International Academy, Research, and Industry Association.
- Wietse de Vries, Martijn Wieling, and Malvina Nissim. 2022. [Make the best of cross-lingual transfer: Evidence from POS tagging with over 100 languages](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7676–7685, Dublin, Ireland. Association for Computational Linguistics.
- Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.
- Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. [Mistral 7b](#). *Preprint*, arXiv:2310.06825.
- Ram Mohan Rao Kadiyala. 2024. Large language models for cross-lingual emotion detection. In *Proceedings of the 14th Workshop of on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis@ACL 2024*, Bangkok, Thailand.
- Tzu-mi Lin, Zhe-Yu Xu, Jian-Yu Zhou, and Lung-Hao Lee. 2024. Assembling large language models for cross-lingual emotion and trigger detection. In *Proceedings of the 14th Workshop of on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis@ACL 2024*, Bangkok, Thailand.
- Aaron Maladry, Alessandra Teresa Cignarella, Els Lefever, Cynthia van Hee, and Veronique Hoste. 2024. [Human and system perspectives on the expression of irony: An analysis of likelihood labels and rationales](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8372–8382, Torino, Italia. ELRA and ICCL.
- Yann Mathet, Antoine Widl  cher, and Jean-Philippe M  tivier. 2015. The unified and holistic method gamma (γ) for inter-annotator agreement measure and alignment. *Computational Linguistics*, 41(3):437–479.
- Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. [SemEval-2018 task 1: Affect in tweets](#). In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.
- Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. [Crosslingual generalization through multitask finetuning](#). *Preprint*, arXiv:2211.01786.
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leonie Aleman, Diogo Almeida, Janko Altmenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Sim  n Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David

- Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O’Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. [Gpt-4 technical report](#). *Preprint*, arXiv:2303.08774.
- Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. [MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.
- Holly A Schroth, Jon Bain-Chekal, and David F Caldwell. 2005. Sticks and stones may break bones and words can hurt me: Words and phrases that trigger emotions in negotiations and their effects. *International Journal of Conflict Management*, 16(2):102–127.
- Pranaydeep Singh, Aaron Maladry, and Els Lefever. 2023. Annotation Guidelines for Labeling Emotion in Multilingual Tweets: Version 1.0 - LT3 technical report – LT3 23-12.
- Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.
- Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussonot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikula, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitaogong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. [Gemma: Open models based on gemini research and technology](#). *Preprint*, arXiv:2403.08295.
- Jesus Vázquez-Osorio, Gerardo Eugenio Sierra-Martínez, Helena Gómez-Adorno, and Gemma Bel-Enguix. 2024. *Pcicunam at wassa 2024: Cross-lingual emotion detection task with hierarchical classification and weighted loss functions*. In *Proceedings of the 14th Workshop of on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis@ACL 2024*, Bangkok, Thailand.
- Pedro Vázquez-Osorio and Helena Gómez-Adorno. 2024. *Lliimas at smm4h: Zero-shot cross-lingual emotion detection in multilingual tweets*. In *Proceedings of the 14th Workshop of on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis@ACL 2024*, Bangkok, Thailand.
- Yuqi Wang, Zimu Wang, Nijia Han, Wei Wang, Qi Chen, Haiyang ZHANG, Yushan Pan, and Anh Nguyen. 2024. Knowledge distillation from monolingual to multilingual models for intelligent and interpretable multilingual emotion detection. In *Proceedings of the 14th Workshop of on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis@ACL 2024*, Bangkok, Thailand.
- Feng Xiong, Jun Wang, Geng Tu, and Ruifeng Xu. 2024. *Hitsz-hlt at wassa-2024 shared task 2: Language-agnostic multi-task learning for explainability of cross-lingual emotion detection*. In *Proceedings of*

the 14th Workshop of on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis@ACL 2024, Bangkok, Thailand.

Jinghui Zhang, Yuan Zhao, Siqin Zhang, Ruijing Zhao, and Siyu Bao. 2024. Enhancing cross-lingual emotion detection with data augmentation and token-label mapping. In *Proceedings of the 14th Workshop of on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis@ACL 2024*, Bangkok, Thailand.

Banghua Zhu, Evan Frick, Tianhao Wu, Hanlin Zhu, and Jiantao Jiao. 2023. Starling-7b: Improving llm helpfulness harmfulness with rlaif.

Mikel Zubillaga, Oscar Sainz, Ainara Estarrona, Oier Lopez de Lacalle, and Eneko Agirre. 2024. [Event extraction in basque: Typologically motivated cross-lingual transfer-learning analysis](#). *Preprint*, arXiv:2404.06392.

Jakub Šmíd, Pavel Pribá, and Pavel Král. 2024. Uwb at wassa-2024 shared task 2: Cross-lingual emotion detection. In *Proceedings of the 14th Workshop of on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis@ACL 2024*, Bangkok, Thailand.