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 emotion 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

Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 454–463

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.

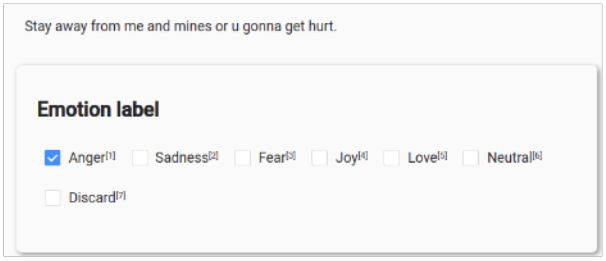

Figure 1: Sub-task 1: Annotation of Emotion label

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.

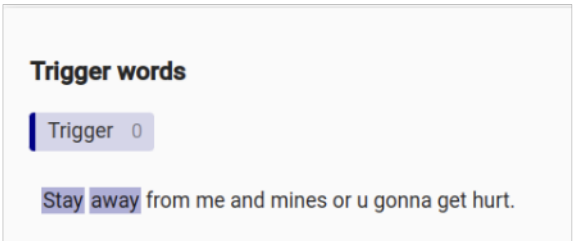

Figure 2: Sub-task 2: Annotation of trigger words.

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 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 multilingual 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

	sad	*about*	*my*	*rejected*	*paper*	*:(*
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

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

Integrated Gradients (Sundararajan et al., 2017) algorithm with transformers-interpret[2] 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

	sad	*about*	*my*	*rejected*	*paper*	*:(*
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 workflows (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-Opus[5], 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

[2] https://github.com/cdpierse/transformers-interpret

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

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

[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
1024m (Kadiyala, 2024)	0.6295
BCSZ (Cheng et al., 2024a)	0.6046
Treehouse (Cheng et al., 2024b)	0.6015
NYCU-NLP (Lin et al., 2024)	0.5951
CTYUN-AI	0.5911
HITSZ-HLT (Xiong et al., 2024)	0.591
UWB (Šmíd et al., 2024)	0.591
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 (Vázquez-Osorio and Gómez-Adorno, 2024)	0.5067
EXALT-Baseline	0.4476

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

Team	Token F1
CTcloud (Zhang et al., 2024)	0.6158
HITSZ-HLT (Xiong et al., 2024)	0.6095
UWB (Šmíd et al., 2024)	0.5919
NLPNewcomer	0.5785
NYCU-NLP (Lin et al., 2024)	0.5636
BCSZ (Cheng et al., 2024a)	0.4778
EXALT-Baseline	0.2349

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.

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