Towards a new Benchmark for Emotion Detection in NLP: A Unifying Framework of Recent Corpora

Anna Koufakou **Elijah Nieves** **John Peller**
Department of Computing & Software Engineering, Florida Gulf Coast University, USA
akoufakou@fgcu.edu

Abstract

Emotion recognition in text is a complex and evolving field that has garnered considerable interest. This paper addresses the pressing need to explore and experiment with new corpora annotated with emotions. We identified several corpora presented since 2018. We restricted this study to English single-labeled data. Nevertheless, the datasets vary in source, domain, topic, emotion types, and distributions. As a basis for benchmarking, we conducted emotion detection experiments by fine-tuning a pretrained model and compared our outcomes with results from the original publications. More importantly, in our efforts to combine existing resources, we created a unified corpus from these diverse datasets and evaluated the impact of training on that corpus versus on the training set for each corpus. Our approach aims to streamline research by offering a unified platform for emotion detection to aid comparisons and benchmarking, addressing a significant gap in the current landscape. Additionally, we present a discussion of related practices and challenges. Our code and dataset information are available at https://github.com/a-koufakou/EmoDetect-Unify. We hope this will enable the NLP community to leverage this unified framework towards a new benchmark in emotion detection.

1 Introduction

Detecting emotions in language, such as *anger, joy, or sadness*, is a powerful application of Natural Language Processing (NLP) with significant interest, especially in recent years (Mohammad et al., 2018; Oberländer and Klinger, 2018; Demszky et al., 2020; Lamprinidis et al., 2021; Plaza-del Arco et al., 2024). Emotion detection is sometimes confused with Sentiment Analysis, a much simpler task that focuses on detecting polarity of sentiments or opinions (Mohammad, 2022). Automated emotion detection is considerably more nuanced and complex due to the subjective and intricate nature of emotions.

NLP-based emotion detection uses datasets annotated with emotions. There is great variability in emotion annotation, including differences in annotation levels (e.g., basic vs. detailed) and labeling schemes (e.g., single vs. multi-label). An even more important challenge is which emotions to use in order to annotate data. Various emotion taxonomies or theories have been presented. Ekman (1992) provided 6 basic emotions: *anger, disgust, fear, joy, sadness, surprise*. Plutchik (1984) proposed a wheel of 8 emotions, adding *trust* and *anticipation* to Ekman's, also presenting dyads (feelings composed of two emotions). Shaver et al. (1987) identified 6 basic emotions: *love, joy, anger, fear, sadness, surprise*, on which they also provided secondary and tertiary levels in a tree-like structure, later refined in Parrott (2001). The *Appraisal* theory (Scherer, 1999; Lazarus, 1991) linked emotions to a persons interpretation of a situation or event. Recently, Cowen and Keltner (2017) identified 27 distinct categories based on videos, facial expressions etc., revised by Demszky et al. (2020) for text-based emotion recognition. Despite the variety of theories available, many efforts have concentrated on single-labeled corpora with a limited set of basic emotions such as Ekman (Plaza-del Arco et al., 2024), likely because these are typically easier for NLP models to handle. Nevertheless, the presence of multiple theories allows different approaches to emotion annotation, making it complicated to unify different datasets for comparisons and benchmarking.

In recent years, numerous emotion-annotated corpora have been introduced from diverse sources and domains, such as social media posts given specific tags or essays on specific topics. Any such available corpora are found in separate repositories and articles, making it challenging for re-

Proceedings of the 2nd GenBench Workshop on Generalisation (Benchmarking) in NLP, pages 196–206
November 16, 2024 ©2024 Association for Computational Linguistics

searchers to be aware of all available resources in order to fully investigate their use. It is noteworthy that a few corpora are well-known, e.g., GoEmotions (Demszky et al., 2020) or TweetEval (Barbieri et al., 2020) with more than 700 citations each, in contrast to others such as Github-love (Imran et al., 2022) with around 20 citations.[1]

As a result, many studies have relied on a subset of available resources, and there has been limited work towards benchmarking. The work by Oberländer and Klinger (2018) stands out: they analyzed and aggregated 14 popular emotion-annotated corpora into a unified framework, in 2018. They used their unified corpus for benchmark results with in-corpus and cross-corpus experiments. The unified framework available online[2] facilitated comparisons of the different corpora. Recent surveys (Alswaidan and Menai, 2020; Acheampong et al., 2020; Nandwani and Verma, 2021; Deng and Ren, 2023; Kusal et al., 2023) do not cover many of the corpora we present in this paper. Very recently, an excellent recent paper by Plaza-del Arco et al. (2024) reviewed over 150 ACL papers (2014-2022), and offered a detailed overview of practices, gaps, and guidelines for emotion analysis in text. Still, their paper did not provide a unified framework or experimentation results.

To address this gap, our paper introduces a unifying framework of text corpora annotated with emotions, as presented in the literature since 2018. We chose 2018 because: (a) it was the most recent year when a unifying framework was presented (Oberländer and Klinger, 2018), and (b) it marked a notable increase in the number of related studies (Plaza-del Arco et al., 2024). Specifically, we identified 11 publicly available emotion-annotated text corpora: we focused our experimentation to English and single-labeled data.[3] While this may seem limited, it serves as a good representation for a significant portion of existing datasets in this area (Plaza-del Arco et al., 2024). More importantly, our primary aim is to explore how to unify (combine) various datasets annotated with different emotions into a single framework, which is not as straightforward.

We conducted classification experiments with these datasets, while comparing our results with the results reported in the original articles. Based on these corpora, we introduce a unified corpus built by mapping original emotions in the corpora to a common set of emotions. Finally, we present baseline benchmarking results for emotion classification with our unified corpus. The ultimate goal is to aid researchers in the field of text-based emotion recognition by providing a unified resource built on a comprehensive set of recent data, which they can access in one repository. Our secondary goal is to furnish a classification baseline benchmark with valuable insights they can use while conducting their own experiments.

The following sections provide descriptions of the corpora (Section 2), details of the unified corpus we created (Section 3), results of our emotion classification experiments (Section 4), and a discussion of findings and observations (Section 5), followed by our concluding remarks and future research directions (Section 6).

2 Corpora

Table 1 summarizes the corpora used in this paper. Table 2 shows which emotions are represented in each corpus. In the following, we provide a brief description of each dataset. We then provide an overview of the datasets and their characteristics. We renamed certain datasets due to unclear or long names.

CARER Saravia et al. (2018) collected tweets with a set of hashtags they constructed, e.g. #depressed, #grief for *sadness*, or #fear, #worried for *fear*. These hashtags were used to annotate the data (*distant supervision*). The dataset posted on Hugging Face is labeled with Shaver, and it is a variant of the dataset presented in the article.

Covid-worry This dataset contains survey responses collected in UK over 2020-22, starting with the first COVID-19 lockdown (Kleinberg et al., 2020). Participants wrote short and long texts, along with demographic data and self-ratings for several emotions. They also chose one emotion among *anger, anxiety, disgust, desire, fear, happiness, relaxation, sadness*. In 2023, the authors presented a 3-year dataset (van der Vegt and Kleinberg, 2023).

EmoEvent Plaza-del-Arco et al. (2020) collected tweets related to events in 2019, and then followed certain steps to select a subset of *affective* tweets. The resulting tweets in English and

[1] Per Google scholar, last checked Aug. 2024.

[2] `https://github.com/sarnthil/unify-emotion-datasets`

[3] Our work uses GoEmotions, which is multi-labeled. We transformed it into single-labeled for this study.

dataset	**source**	**# emotions**	**size**	**reference**	**avail.**
CARER	tweets	6	417*	(Saravia et al., 2018)	HG
Covid-worry	essays	8	5.2	(van der Vegt and Kleinberg, 2023)	G,O
EmoEvent	tweets	6+1	7.3	(Plaza-del-Arco et al., 2020)	G
enISEAR	self-written	8	1	(Troiano et al., 2019)	O
Github-love	github	6	1.7	(Imran et al., 2022)	HG
GoEmotions	reddit	27+1, 6+1	58*	(Demszky et al., 2020)	G
GoodNews	headlines	15+1	5	(Oberländer et al., 2020)	Uni
StackOv-GS	stack overflow	6	4.8	(Novielli et al., 2018)	G
TweetEval	tweets	4	5	(Barbieri et al., 2020)	G
Universal Joy	facebook	5	284*	(Lamprinidis et al., 2021)	G
WASSA-21	essays	6+1	2.6	(Tafreshi et al., 2021)	Cd, Rq

Table 1: Summary of datasets used in this paper. Size in thousands (rounded to the closest hundred; if corpus has multiple languages, it refers to English; * denotes that we used a smaller sample of this dataset for our experiments). '+1' in '# emotions' column denotes additional class for *neutral/no emotion/other(s)*. 'avail.' is data availability: Cd=Codalab, G=Github, HG=Hugging Face, Kg=Kaggle, O=Other, Rq=By Request (the URLs are provided in our online repository).

in Spanish were annotated by Amazon MTurkers using Ekman plus *other*.

enISEAR Troiano et al. (2019) provided German (deISEAR) and English (enISEAR) corpora, using a framework similar to earlier ISEAR (International Survey on Emotion Antecedents and Reactions) (Scherer and Wallbott, 1994). A questionnaire instructed annotators (by crowdsourcing) to give a description of an event for which they felt a particular emotion. Each record was annotated with Ekman plus *guilt* and *shame*.

Github-love Imran et al. (2022) collected GitHub comments on pull requests/issues for popular repositories, annotated by the authors using Shaver. Besides these basic emotions, they also used detailed levels of emotions (Shaver et al., 1987), where they added some of the emotions presented by Demszky et al. (2020), e.g. *approval* or *confusion*. Note that the dataset available online has basic emotion labels, not the detailed levels in the paper.

GoEmotions Demszky et al. (2020) collected Reddit comments with crowdsourced annotations for 27 emotions or *neutral*, revised from Cowen and Keltner (2017). They also provided an Ekman mapping from their detailed emotions. This dataset is multi-labeled and we transformed it into single-labeled for the purposes of this study (see Section 3 for details).

GoodNews Oberländer et al. (2020) collected English news headlines and annotated them via crowdsourcing (named GoodNewsEveryone in the original article). Annotations were provided for emotions (extended Plutchik) and their intensity, as well as semantic roles (such as experiencer or cause), and reader interpretation of the headline.

StackOv-GS Novielli et al. (2018) collected Stack Overflow questions, answers and comments for the 'StackOverflow Gold Standard'. They were annotated by volunteers with Shaver.

TweetEval Barbieri et al. (2020) created a unified twitter dataset with seven heterogeneous Twitter-specific classification tasks. Among those, they included Affect in Tweets (Mohammad et al., 2018) only keeping single-label records and dropping rare emotions. This resulted in records labeled with *anger, joy, optimism, sadness*.

Universal Joy Lamprinidis et al. (2021) presented a dataset with anonymized public Facebook posts that were originally collected in 2014 in 18 languages. The authors labeled the records with *anger, anticipation, fear, joy, sadness*.

WASSA-21 This dataset was part of a shared task in the 11th *Workshop on Computational Approaches to Subjectivity, Sentiment & Social Detection and Emotion Classification (WASSA)*, summarized by Tafreshi et al. (2021). It contains essays written to express the authors empathy and distress in reaction to news articles related to harm. The emotion labels (Ekman) were first predicted by Neural Networks and then post-annotated by crowdsourcing workers and a PhD student.

Overview Out of the 11 datasets in Table 1, 5 came from social media (X/Twitter, Facebook, Reddit), 2 came from software-related websites (GitHub and Stack Overflow), 1 was with news

dataset	annotation	A	Ant	D	F	J	Ne	Sa	Su	T	other emotions
CARER	Shaver	✓	–	–	✓	✓	–	✓	✓	–	love
Covid-worry	Other	✓	–	✓	✓	✓	–	✓	–	–	anxiety, desire, relaxation
EmoEvent	E+Ne	✓	–	✓	✓	✓	✓	✓	✓	–	
enISEAR	Ext. E	✓	–	✓	✓	✓	–	✓	–	–	shame, guilt
Github-love	Shaver	✓	–	–	✓	✓	–	✓	✓	–	love
GoEmotions	E+Ne, Revised CK	✓	–	✓	✓	✓	✓	✓	✓	–	27 fine-grained emotions
GoodNews	Ext. P	✓	✓	✓	✓	✓	-	✓	✓	✓	annoyance, guilt, love, pride, shame
StackOv-GS	Shaver	✓	–	–	✓	✓	-	✓	✓	–	love
Tweeteval	Other	✓	–	–	–	✓	–	✓	–	–	optimism
Universal Joy	Mod. P	✓	✓	–	✓	✓	–	✓	–	–	
WASSA-21	E+Ne	✓	–	✓	✓	✓	✓	✓	✓	–	
	Total	11	2	6	10	11	3	11	7	1	

Table 2: The emotions in each corpus. E=Ekman, P=Plutchik, and CK=Cowen & Keltner; Mod.=modified, Ext.=extended. "–" means the emotion is not in that corpus. Emotions: A-Anger, Ant-Anticipation, D-Disgust, F-Fear, J-Joy, Ne-Neutral or no emotion or other, Sa-Sadness, Su-Surprise, and T-Trust.

headlines, and 3 were self-reported (for example, self-written responses to questions and self-ratings of emotions in Covid-worry). The data that were based on online posts or comments were usually annotated by humans (volunteers, experts or crowdsourcing workers), though CARER used the hashtags as noisy labels. The self-reported varied: in Covid-worry, essays were written by survey participants related to their current situation, while in enISEAR the statements were written by crowdsourcing workers: they were given an emotion, and were asked to describe a related event.

As far as size, most corpora are small; the two smallest are enISEAR and Github-love. There are 3 larger datasets: CARER and Universal Joy (hundreds of thousands) and GoEmotions (about 58K). Finally, all corpora follow basic emotion annotation, except GoEmotions and GoodNews.

Based on Table 2, we observed that all or most corpora contain *anger*, *fear*, *joy*, and *sadness*; frequently represented emotions are *disgust* and *surprise*; *love* is represented in fewer than half of the corpora; the *anticipation* and *trust* emotions followed by *neutral, no emotion, other* are the least represented in the data.

Finally, the distribution of emotions varies across the corpora. Some corpora exhibit a range of dominant emotions versus very low representation of certain emotions. For instance, CARER is primarily dominated by *joy*, followed by *sadness*, and it has a very low sample of *surprise*. Universal Joy is heavily dominated with *anticipation* and then *joy*, while low on *anger* and *fear*. Covid-worry is led by *anxiety* and *fear*, with *joy* trailing behind, and it has a very low number of records with *disgust* or *anger*. StackOV-GS is led by *love*, followed by *anger*. EmoEvent is predominantly *neutral* ('*other*'), followed by *joy*. Both EmoEvent and StackOV-GS are very low in *fear* and *surprise*. Finally, *disgust* has very low representation in most datasets. As an exception, enISEAR is balanced as crowdsourcing workers were asked to write a certain number of statements for each of the emotions.

3 Creating a Unified Corpus

First, for any corpus we downloaded, we spent effort reading instructions and exploring file formats, features, labeling schemes, etc. For example, some data had integers as labels, which we had to map to emotions per author instructions; some data came with many features so we had to extract text/labels. Many sets were well-organized and documented, with a couple of exceptions that were harder to understand and transform. In short, we spent significant effort to integrate diverse corpora into the unified corpus with the goal to save other researchers time and effort.

Based on the emotions in Table 2 and previous work (Oberländer and Klinger, 2018; Demszky et al., 2020), we defined a scheme roughly following Plutchik and Shaver as our common emotion label set. Specifically, we used *anger, anticipation,*

unified	original
anger	anger, annoyance, annoyed, shame
anticipation	anticipation, neg. or pos. anticipation
disgust	disgust
fear	fear, anxiety
joy	joy, happiness, happy, desire, optimism, optimistic, pride, relaxation
love	love, love incl. like
neutral	neutral, none, noemo, other
sadness	sadness, sad, guilt
surprise	surprise, neg. surprise, pos. surprise
trust	trust

Table 3: The mapping we followed for mapping original labels to unified labels. We roughly follow the models by Plutchik (1984) and Shaver et al. (1987).

disgust, fear, joy, love, sadness, surprise, trust, and *neutral*; we kept *neutral* due to its relatively good representation in certain corpora (e.g., it was about 33% of the records in GoEmotion).

We mapped original emotions from the data to the emotions in the common set as shown in Table 3. We decided on the mappings based on previous literature, and by observing sample records from each corpus. For example, Oberländer and Klinger (2018) used a very similar set: their list had the same emotions as ours except they included *confusion* and not *love*. Also, Demszky et al. (2020) mapped *annoyance* to *anger* and *optimism* and *pride* to *joy*. For GoEmotions (a multi-labeled corpus with detailed emotions), we kept only records with a single label or if the multiple labels mapped to the same label in our common set of emotions (note that the creators provided Ekman mappings of their detailed labels). This resulted in a dataset with 43,975 records. For Covid-worry, we combined all surveys from 3 years. Finally, due to our resources, we downsampled CARER and Universal Joy to a more manageable size for our experiments, keeping emotion distribution the same as in the original corpora. As a result, in our experiment there were 62,522 records for CARER, and 84,695 for Universal Joy (about 30% of the original size).

3.1 Unified Corpus Properties

Our unified corpus addresses the following properties important for generalization testing as shown by Hupkes et al. (2023). In all the points below we refer the reader to the dataset descriptions in the earlier sections and Tables 1 and 2.

Platform Shift: The datasets that were collected from online sources were sourced from different platforms: Twitter, Reddit, Facebook, Github, StackOverflow.

Language Shift: Even though most datasets came from social media or online forums, there were also datasets that contain self-written statements or news headlines.

Topic Shift: The datasets were collected for different reasons and topics, for example, Covid-19 (Covid-worry), events (EmoEvent), or software (code) questions and comments (StackOv-GS).

Emotion Shift: The emotions represented in each corpus as well as their distributions vary, for example some corpora are heavily dominated by positive emotions (CARER or Universal Joy), while others by negative emotions (Covid-worry or WASSA-21).

4 Experiments and Results

4.1 Experimental Setup

We used Google colab[4] to run all our experiments. For our classification experiments, we selected to use `distilroberta-base`:[5] it is a distilled version of RoBERTa (Robustly optimized BERT approach) (Liu et al., 2019), with 6 layers, 768 dimension and 12 heads, resulting in a total of 82M parameters (compared to 125M parameters for RoBERTa-base). We fine-tuned the model for 2 Epochs, with learning rate of $1e^{-5}$, maxlen of 256 and batch of 8, based on early trials. If the corpus came with a train/test set (e.g., WASSA-21), we used those sets, otherwise we used an 80-20 stratified split. We repeated each experiment 5 times and reported the average f1-score. In total, we performed 110 experiments, either fine-tuning the model on each single corpus (5 runs × 11 corpora = 55 total experiments, see results in Section 4.2 and 4.3), or fine-tuning on the Unified train set (also 55 total, see results in Section 4.4). We also performed some additional cross-corpus experiments as examples (see Section 4.5).

[4] https://colab.research.google.com/

[5] https://huggingface.co/distilroberta-base

Corpus	Ours	Previous work (OA = Original Article)
CARER	91%	OA used larger and/or different version of the corpus, different models: max f1-macro 79%.
Covid-worry	46%	No previous work used combined data from all 3 surveys.
EmoEvent	34%	OA used SVM with 32% f1-macro.
enISEAR	48%	OA used MaxEnt with 47% f1-micro.
Github-love	44%	OA used various models (non-transformers) with max f1-macro of 44%.
GoEmotions	65%	OA used BERT-base with 64% f1-macro on their Ekman taxonomy version, 46% on data with their own taxonomy.
GoodNews	26%	We could not find previous results for emotion detection.
StackOv-GS	44%	We could not find previous results for emotion detection.
TweetEval	80%	OA used RoBERTa-base with 76% f1-macro.
Universal joy	63%	OA had multiple-sized data, ours is downsampled. Their mBERT results showed 46-63% f1-macro.
WASSA-21	31%	Results for shared task in OA ranged in 31-55%. Top ranking teams used ensembles/augmented with GoEmotions.

Table 4: A comparison of our f1-macro results for each corpus in our benchmark versus existing results from the original literature on their datasets. See Table 1 for references (Original Article) related to each corpus.

4.2 Overall Performance and Comparison with Previous Work

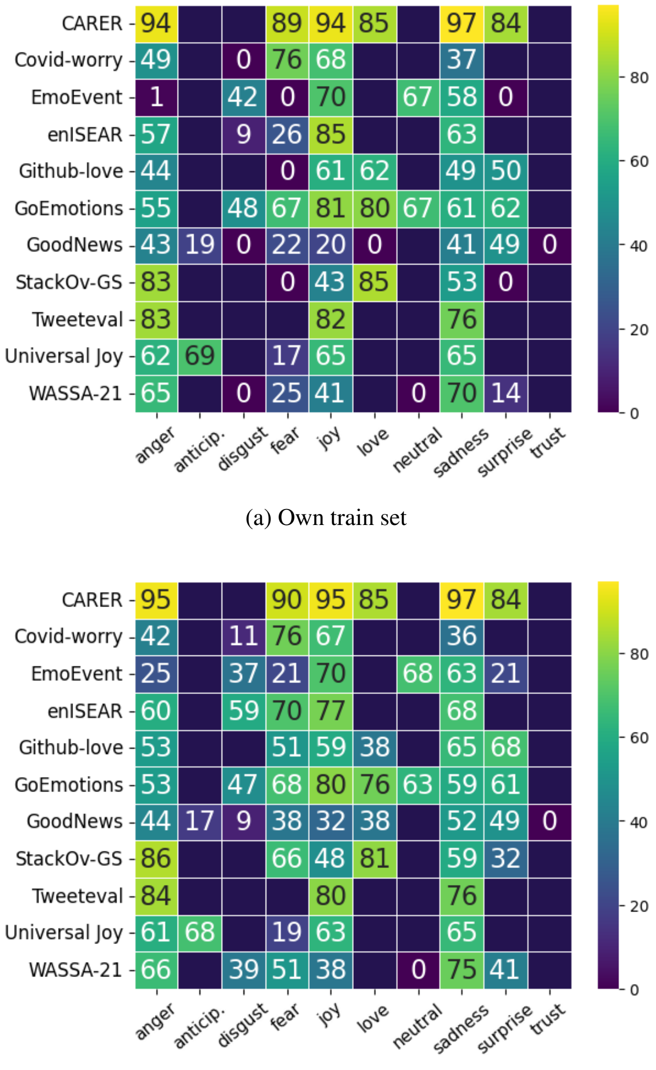
We first show our f1 macro-averaged over all emotions for each corpus versus the results for each specific corpus as shown in the related literature. The reader should keep in mind that there were differences between some of the datasets we used in this work versus the ones in previous work in the literature: for example, we downsampled very large datasets such as CARER, and we combined 3 surveys in Covid-worry (the same survey was given in 3 consecutive years, see section 2). Also, the experimental setup (such as train-test split or the (hyper-)parameters of a model) in existing work usually varies from our work. Nevertheless, we understand that such a comparison might be beneficial to show a ‘bigger’ picture for the reader. Therefore, we provide comparative results in Table 4. Finally, even though we concentrated on original work that presented the datasets, if that work did not have emotion detection results, we also looked in the recent literature. For example, we previously applied RoBERTa-based on the first survey from Covid-worry resulting in 49% f1-macro (Koufakou et al., 2022), but, to our knowledge, no previous work has used all 3 surveys.

Overall, CARER had the best performance (91% f1-macro), followed by TweetEval (80% f1-macro), Universal Joy and GoEmotions (f1-macros in the 60’s). The rest of the datasets had f1-macro values ranging from high 40’s to low 30’s. Our baseline’s f1-macros are largely similar to previous literature results. In certain cases, our results are lower than the literature, e.g. for WASSA-21: the top ranking teams in that shared task augmented the train set with GoEmotions, and usually also employed ensembles of models. In other cases, our f1-macros are higher, e.g. for CARER: they used a different version of their dataset in their article as opposed to the one publicly shared.

4.3 Results per Emotion

Figure 1a depicts the f1-score per emotion for each corpus as a heatmap. Per emotion, CARER had the best f1-score, except for the emotions it did not contain (*disgust* and *neutral*): GoEmotion had the best f1-score for those.

Looking at specific emotions, the hardest emotions to detect were *disgust*, *fear*, *surprise*, and *trust*, depending on the dataset. First, as also observed by Oberländer and Klinger (2018), emotions with low frequency were harder to detect. As an example, in Covid-worry, *disgust* had the lowest frequency by far. When we inspected a resulting confusion matrix, *disgust* was mostly confused with *anger* and *fear*. In other sets, rarest emotions were mispredicted completely: *disgust* and *neutral* in WASSA-21, *fear* and *surprise* in StackOv-GS and in EmoEvent. Several of these corpora are imbalanced and have been shown to benefit from techniques such as data augmentation. The winner in WASSA-21 showed that augmenting their training with GoEmotions improved classification



(a) Own train set

(b) Unified train set

Figure 1: Heatmaps with f1-score per emotion (x-axis) for each corpus (y-axis) with fine-tuning either (a) using the train set of the corpus or (b) using the Unified train set. Empty cell: dataset does not contain that emotion.

(Mundra et al., 2021). In that vein, we show the results of fine-tuning the model with the Unified train set in section 4.4, and a few cross-corpus experiments in Section 4.5.

Besides imbalanced distribution, annotation of the emotions plays a role. For example, Plaza-del-Arco et al. (2020) observed that annotators for EmoEvent had trouble with *fear*, *disgust* and *surprise*, and distinguishing between *anger* and *disgust* (complementary emotions). In Github-love, *joy* was mispredicted many times as *love*. We examined random comments and found that some were so similar that a human would struggle to distinguish between them: e.g., “This will answer your question: Good luck!" (*love*) and “excellent, good luck!" (*joy*).

4.4 Results from Fine-Tuning on the Unified Train Set

First, we created a ‘Unified train’ set from merging all train sets from all unified corpora: this was after each corpus had been transformed to the same format and our common label set. This results in a train set of about 180.6K records. We observed that the Unified train set is heavily skewed towards *joy* (about 34%), then *sadness* and *anticipation* (about 16% each). These emotions are heavily represented in the larger sets (CARER, Universal Joy, then GoEmotions). For the experiments in this Section, we fine-tuned the model on that Unified train set and predicted the labels of the test set from each corpus. The overall results (f1-macro) from fine-tuning the model on this Unified train

corpus	own	unified	Δ
CARER	91%	91%	0%
Covid-worry	46%	46%	0%
EmoEvent	34%	44%	**10%**
enIsear	48%	67%	**19%**
Github-love	44%	56%	**12%**
GoEmotion	65%	64%	-1%
GoodNews	22%	31%	**9%**
StackOv-GS	44%	62%	**18%**
TweetEval	80%	80%	0%
Universal Joy	64%	63%	-1%
WASSA-21	31%	44%	**13%**

Table 5: f1-macro after fine-tuning the model on its own train set versus on the Unified train set, followed by the difference (Δ). Both were tested on the same test set. Bold: improvement larger than 5%.

set versus only on the original train set from each corpus are shown in Table 5.

To summarize, CARER, Covid-Worry, GoEmotions, TweetEval and Universal Joy did not show an improvement when training on the Unified train versus just training on the original train set. Most of these datasets already had highest results (see Fig. 1a and Table 4). For example, CARER was a large dataset with 91% f1-macro, so there was little room for improvement. However, Covid-worry still had low f1-macro in Table 5: our observation is that corpus is largely dominated by *worry* and *anxiety* which is not represented well in the other corpora. We did map *anxiety* to *fear* in order to create the Unified corpus; still, the emotions in Covid-worry do not seem to translate well to the ones in the rest of the corpora.

On the other hand, 6 out of 12 corpora showed improvements ranging in 9-19% (see the Δ column in Table 5). Specifically, the f1-macro improvement was around 10% for 4 corpora (EmoEvent 10%, Github-love 12%, GoodNews 9%, and WASSA-21 13%) and about 20% for two corpora (enISEAR 19% and StackOv-GS 18%).

We can also look at specific emotions shown as a heatmap in Figure 1b. For example, in StackOv-GS, *fear* and *surprise* were not detected at all (0% for own train set in Fig. 1a) versus f1-scores of 66% and 32% respectively (Unified train in Fig. 1b). Overall, we observed from the two heatmaps, there were improvements for *disgust*, *fear* and *surprise*, which were either relatively rare or they overlapped, as discussed in earlier sections.

4.5 Additional Experiments

Due to our constraints of time and resources, we were not able to conduct a full cross-corpus experimentation. This could mainly consist of training on the train set of one corpus and then testing on the test set of another corpus or, following (Oberländer and Klinger, 2018), training on one (entire) corpus and evaluating on a different (entire) corpus. Nevertheless, we performed some initial cross-corpus experiments, briefly summarized here as potential ideas for this unified resource. The reader is directed to Table 4 for f1-macro results when training on each original train set, for comparison purposes.

For instance, one could explore the effect of data source. As an example, we trained on the GoEmotion train set (social media posts), then tested on the EmoEvent test set (also social media posts) and on the WASSA-21 test set (self-written essays). The f1-macro for EmoEvent was 34%, which matched the results based on its own train set. The f1-macro for WASSA-21 was 38%, better than training on its own train set by 7%. A challenge in the cross-domain setup is handling the differences in emotion labels across datasets which are combined in these experiments (also true for the Unified train in Section 4.4).

As an example of exploring datasets with matching original emotion labels, we trained on the CARER train set and tested on GitHub-love and StackOv-GS (all featuring Shaver emotions originally). Both tests yielded f1-macro in the mid-30s, compared to mid-40s when training on their respective train sets. It is noteworthy that CARER consists of tweets, while the other two datasets have code-related comments. Also, the most frequent emotion for all three datasets is positive (*joy* or *love*), but the second most frequent emotion in CARER is *sadness* (29%), versus *anger* in GitHub-love and StackOv-GS (20-30% depending on the dataset).

5 Discussion

In this paper, we started by selecting 11 recently introduced datasets with emotion-annotated records in order to introduce a new unified framework for benchmarking emotion detection in NLP. We described the characteristics of these datasets, which vary in size, topic, source, emotions and distributions. Nevertheless, one could question the selection of the specific datasets. It would be beneficial to explore earlier datasets or datasets we did not include, and how they compare/connect to corpora in this work: we leave this for future work. Moreover, we focused on English corpora, unfortunately, a common limitation in NLP: a multilingual study is thus needed, e.g. see a multilingual sentiment analysis study by Rajda et al. (2022). Specifically for emotion detection, one should consider linguistic and cultural differences for emotions (De Bruyne, 2023).

We combined the datasets into one unified framework by mapping to a common set of emotions. We followed a simple emotion scheme for this: each record gets assigned a single label out of n emotions, similar to earlier work (Oberländer and Klinger, 2018). Many of the available corpora only have a few basic emotions to start with (Oberländer and Klinger, 2018; Plaza-del Arco et al., 2024). In reality, though, emotions are complex and an individual's writings may encapsulate multiple emotions. This is even more prevalent in essays with multiple sentences (Tafreshi et al., 2021). To more accurately reflect human emotions, fine-grained emotion annotations are often preferred over coarse ones (Demszky et al., 2020).

Our review of the datasets and related literature revealed several limitations and issues similar to those identified in previous work, such as inconsistencies in annotation practices or inadequate reporting of the annotation process (Stajner, 2021; Plaza-del Arco et al., 2024). These issues underscore the need for a more thorough analysis of practices and benchmarking within available corpora. Good examples in other areas include a systematic review in hate speech detection by Poletto et al. (2021). Also, we found research in generalized offensive language identification by Dmonte et al. (2024): they used relatively basic labels, e.g. *Offensive* or *Non-offensive*. In contrast, as we have shown in this paper, emotion-annotated data involves more complexity and thus additional challenges. Additionally, automated emotion recognition carries ethical considerations as shown by Mohammad (2022), who proposed an ethics sheet outlining 50 ethical considerations. For instance, the need to account for both the speaker's and the reader's perspectives, which can vary significantly from one individual to another.

While this paper represents considerable effort, there is still more work to match the underlying complexity of this study. Our classification experiments offer a baseline rather than a complete benchmark. Our published code and list of dataset links will enable anyone to recreate and further utilize the unified , and even possibly extend it. The research community is welcome to use these resources and employ various models and/or explore the effect of different hyper-parameters on the results.

6 Conclusions

This paper answers the imperative need for studying recent text-based corpora annotated for emotion detection. Our investigation into diverse corpora sourced from various domains and introduced since 2018 summarizes and gives insights into their characteristics, such as source, topic, size, emotion and distributions, etc. Furthermore, we constructed a unified framework built from these corpora by mapping their emotions to a common set of labels. We used these resources to conduct emotion detection experiments, and compared the effect of fine-tuning a pretrained model to the train set of each corpus versus to the unified train set. This consolidated platform will be a valuable resource for researchers, streamlining efforts and providing the basis for a practical emotion classification benchmark.

While this paper represents considerable effort, there is still more work to match the underlying complexity of this study. Future directions include expanding this work to additional datasets, including multi-lingual and multi-label settings, while also conducting additional experiments (e.g. cross-corpus or various classifiers) and delving deeper into annotation practices and methodologies.

7 Limitations and Ethical Considerations

This work included datasets presented since 2018, all in English and all represented as single-labeled. As we discussed in Section 5, the dataset selection for benchmarking should be more expansive, not only in terms of languages and emotion labeling but also regarding data sources and topics. In the realm of emotion recognition using NLP, the linguistic and cultural diversity of emotions highlights the need for more inclusive and representative datasets. Furthermore, additional experimentation, especially cross-corpus, is essential to establish the unified framework as a valuable benchmark.

Our work did not collect or annotate any datasets, and instead used publicly available datasets. Nevertheless, it is important that any such research in the field of automated emotion recognition, including the curation of emotion annotated datasets, should consider ethical questions such as the ones by Mohammad (2022).

References

Francisca Adoma Acheampong, Chen Wenyu, and Henry Nunoo-Mensah. 2020. Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports*, 2(7):e12189.

Nourah Alswaidan and Mohamed El Bachir Menai. 2020. A survey of state-of-the-art approaches for emotion recognition in text. *Knowledge and Information Systems*, 62(8):2937–2987.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online.

Alan S Cowen and Dacher Keltner. 2017. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the national academy of sciences*, 114(38):E7900–E7909.

Luna De Bruyne. 2023. The paradox of multilingual emotion detection. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 458–466. Association for Computational Linguistics.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054.

Jiawen Deng and Fuji Ren. 2023. A survey of textual emotion recognition and its challenges. *IEEE Transactions on Affective Computing*, 14(1):49–67.

Alphaeus Dmonte, Tejas Arya, Tharindu Ranasinghe, and Marcos Zampieri. 2024. Towards generalized offensive language identification. *arXiv preprint arXiv:2407.18738*.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Dieuwke Hupkes, Mario Giulianelli, Verna Dankers, Mikel Artetxe, Yanai Elazar, Tiago Pimentel, Christos Christodoulopoulos, Karim Lasri, Naomi Saphra, Arabella Sinclair, et al. 2023. A taxonomy and review of generalization research in nlp. *Nature Machine Intelligence*, 5(10):1161–1174.

Mia Mohammad Imran, Yashasvi Jain, Preetha Chatterjee, and Kostadin Damevski. 2022. Data augmentation for improving emotion recognition in software engineering communication. In *Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering*, pages 1–13.

Bennett Kleinberg, Isabelle van der Vegt, and Maximilian Mozes. 2020. Measuring emotions in the covid-19 real world worry dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*.

Anna Koufakou, Jairo Garciga, Adam Paul, Joseph Morelli, and Christopher Frank. 2022. Automatically classifying emotions based on text: A comparative exploration of different datasets. In *34th International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE.

Sheetal Kusal, Shruti Patil, Jyoti Choudrie, Ketan Kotecha, Deepali Vora, and Ilias Pappas. 2023. A systematic review of applications of natural language processing and future challenges with special emphasis in text-based emotion detection. *Artificial Intelligence Review*, pages 1–87.

Sotiris Lamprinidis, Federico Bianchi, Daniel Hardt, and Dirk Hovy. 2021. Universal joy a data set and results for classifying emotions across languages. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 62–75, Online.

Richard S Lazarus. 1991. Progress on a cognitive-motivational-relational theory of emotion. *American psychologist*, 46(8):819.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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.

Saif M Mohammad. 2022. Ethics sheet for automatic emotion recognition and sentiment analysis. *Computational Linguistics*, 48(2):239–278.

Jay Mundra, Rohan Gupta, and Sagnik Mukherjee. 2021. WASSA@IITK at WASSA 2021: Multi-task learning and transformer finetuning for emotion classification and empathy prediction. In *Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 112–116, Online. ACL.

Pansy Nandwani and Rupali Verma. 2021. A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining*, 11(1):1–19.

Nicole Novielli, Fabio Calefato, and Filippo Lanubile. 2018. A gold standard for emotion annotation in stack overflow. In *Proceedings of the 15th international conference on mining software repositories*, pages 14–17.

Laura Ana Maria Oberländer, Evgeny Kim, and Roman Klinger. 2020. Goodnewseveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1554–1566.

Laura Ana Maria Oberländer and Roman Klinger. 2018. An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th international conference on computational linguistics*, pages 2104–2119.

W Gerrod Parrott. 2001. *Emotions in social psychology: Essential readings*. psychology press.

Flor Miriam Plaza-del Arco, Alba A. Cercas Curry, Amanda Cercas Curry, and Dirk Hovy. 2024. Emotion analysis in NLP: Trends, gaps and roadmap for future directions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5696–5710, Torino, Italia. ELRA and ICCL.

Flor Miriam Plaza-del-Arco, Carlo Strapparava, L. Alfonso Urena-Lopez, and M. Teresa Martin-Valdivia. 2020. EmoEvent: A Multilingual Emotion Corpus based on different Events. In *Proceedings of the 12th Language Resources and Evaluation Conference*.

Robert Plutchik. 1984. Emotions: A general psychoevolutionary theory. *Approaches to emotion*, 1984(197-219):2–4.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55:477–523.

Krzysztof Rajda, Lukasz Augustyniak, Piotr Gramacki, Marcin Gruza, Szymon Woźniak, and Tomasz Kajdanowicz. 2022. Assessment of massively multilingual sentiment classifiers. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 125–140, Dublin, Ireland.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697.

Klaus R Scherer. 1999. *Appraisal theory*. John Wiley & Sons Ltd.

Klaus R Scherer and Harald G Wallbott. 1994. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of personality and social psychology*, 66(2):310.

Phillip Shaver, Judith Schwartz, Donald Kirson, and Cary O'connor. 1987. Emotion knowledge: further exploration of a prototype approach. *Journal of personality and social psychology*, 52(6):1061.

Sanja Stajner. 2021. Exploring reliability of gold labels for emotion detection in Twitter. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1350–1359, Held Online. INCOMA Ltd.

Shabnam Tafreshi, Orphée De Clercq, Valentin Barriere, Sven Buechel, João Sedoc, and Alexandra Balahur. 2021. Wassa 2021 shared task: Predicting empathy and emotion in reaction to news stories. In *Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Online. ACL.

Enrica Troiano, Sebastian Padó, and Roman Klinger. 2019. Crowdsourcing and validating event-focused emotion corpora for german and english. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4005–4011.

Isabelle van der Vegt and Bennett Kleinberg. 2023. A multi-modal panel dataset to understand the psychological impact of the pandemic. *Scientific Data*, 10(1):537.