SRL4E – Semantic Role Labeling for Emotions: A Unified Evaluation Framework

Cesare Campagnano[1] and **Simone Conia**[1] and **Roberto Navigli**[2]
Sapienza NLP Group
[1]Department of Computer Science
[2]Department of Computer, Control and Management Engineering
Sapienza University of Rome
{campagnano,conia}@di.uniroma1.it navigli@diag.uniroma1.it

Abstract

In the field of sentiment analysis, several studies have highlighted that a single sentence may express multiple, sometimes contrasting, sentiments and emotions, each with its own experiencer, target and/or cause. To this end, over the past few years researchers have started to collect and annotate data manually, in order to investigate the capabilities of automatic systems not only to distinguish between emotions, but also to capture their semantic constituents. However, currently available gold datasets are heterogeneous in size, domain, format, splits, emotion categories and role labels, making comparisons across different works difficult and hampering progress in the area. In this paper, we tackle this issue and present a unified evaluation framework focused on Semantic Role Labeling for Emotions (SRL4E), in which we unify several datasets tagged with emotions and semantic roles by using a common labeling scheme. We use SRL4E as a benchmark to evaluate how modern pretrained language models perform and analyze where we currently stand in this task, hoping to provide the tools to facilitate studies in this complex area.

1 Introduction

Emotion detection – a long-standing open problem in Natural Language Processing (NLP) – is the task of automatically associating one or more emotions with a text. Even though emotional states are highly subjective and often depend on several factors, such as one's past experiences, culture and education, the automatic identification, categorization and analysis of emotions in texts has been found to be beneficial in a wide array of downstream tasks, such as hate speech detection (Markov et al., 2021), sarcasm detection (Chauhan et al., 2020), and modeling political discourse (Huguet Cabot et al., 2021), *inter alia*.

In the past decade, Deep Learning techniques have become ubiquitous in the development of automatic systems for an increasing number of NLP tasks, including emotion detection (Chatterjee et al., 2019). However, most of the effective neural-based approaches still require significant amounts of training data in order to learn to perform at their best. For this reason, with a view to bootstrapping the development of neural systems for emotion detection, there have been several efforts to annotate corpora with emotions manually (Bostan and Klinger, 2018).

Nevertheless, over the past few years, numerous studies have indicated that a short text, even a single sentence, may contain multiple – at times concurring, at other times contrasting – sentiments and emotions. And not only this, two emotions in the same sentence may be experienced, targeted, and/or caused by different semantic constituents which, similarly to predicate-argument structures in Semantic Role Labeling (SRL), can be linked to form abstract semantic structures. The potential applications in social media analysis, abuse detection, and other actively studied areas in NLP (Rajamanickam et al., 2020) of such automatically-extracted emotion-focused semantic structures have prompted researchers to create datasets aimed at investigating the capabilities of modern systems to parse emotional events (Oberländer et al., 2020). Unfortunately, despite the increasing interest in this area, currently available gold datasets feature heterogeneous structures and characteristics, ranging from varying sizes to different domains, file format, splits and, most importantly, non-overlapping emotion categories. We argue that this heterogeneity obstructs, or at least hinders, further progress in this relatively new area of sentiment analysis.

In this paper, we take a step towards addressing the above-mentioned issues and introduce a unified framework for Semantic Role Labeling for Emotions (SRL4E). In SRL4E, we unify several gold but heterogeneous datasets that contain anno-

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 4586 - 4601

tations both for emotions and for their semantic constituents, so as to obtain a new homogeneous dataset that covers diverse domains and that can be used to train, validate and evaluate current and future work in this task. Our contributions can be summarized as follows:

- We propose a unified gold benchmark for training and evaluating a system on Semantic Role Labeling for Emotions (SRL4E);
- We take advantage of SRL4E to show the inadequacy of training a model on domain-specific data and the benefits of our unified framework;
- We show the advantages of bilingual transfer from English to Chinese, and vice versa, in SRL4E.

We release SRL4E at `https://github.com/SapienzaNLP/srl4e` in the hope that our unified framework will become a stepping stone for the development and evaluation of current and future approaches to Semantic Role Labeling for Emotions.

2 Related Work

Emotion classification datasets. Currently, there are a wide variety of datasets annotated with emotion classes, ranging across different domains and using different annotation schemes. Among others, we can find datasets on emotional experiences (Scherer and Wallbott, 1994), children's fairy tales (Alm et al., 2005), news headlines (Strapparava and Mihalcea, 2007), blog posts (Aman and Szpakowicz, 2007, 2008), news (Lei et al., 2014), social media posts and reviews (Buechel and Hahn, 2017), dialogs (Li et al., 2017; Chatterjee et al., 2019), Facebook posts (Preoţiuc-Pietro et al., 2016), with many focusing on tweets (Mohammad, 2012; Mohammad and Bravo-Marquez, 2017; CrowdFlower, 2016; Liu et al., 2017; Schuff et al., 2017) due to their tendency to have dense emotional content. To meet such a diversity of contents and formats, Bostan and Klinger (2018) created a unified resource for emotion classification comprising many of the aforementioned datasets, while Tafreshi and Diab (2018), instead, added an additional clause-level annotation layer to some existing resources. More recent efforts, such as GoEmotion (Demszky et al., 2020), XED (Öhman et al., 2020) and CancerEmo (Sosea and Caragea, 2020), provide, respectively, emotion annotations for Reddit comments, multilingual subtitles and blog posts about health problems.

Although the above-mentioned corpora have enabled systems to perform emotion detection across different domains, their annotations are sentence-level and, therefore, introduce an oversimplification: they indicate only the overall sentiment and/or emotion that appears in a given text, neglecting the cases in which a short text, even a single sentence, may express multiple emotions. Furthermore, the aforementioned datasets do not indicate which part of the text elicits an emotion and who experiences, is the target of, or causes that emotion. As a consequence, a system trained on these datasets may produce predictions that are hard to interpret and more difficult to use in real-world applications. To overcome these problems, we rely on resources that not only indicate emotions, but also identify their semantic constituents, namely, emotional CUES, EXPERIENCERS, TARGETS and STIMULI.

Emotion Taxonomy. Among the studies that aim to identify the fundamental emotions, Ekman (1992) proposed a set of six categories: *anger*, *disgust*, *fear*, *joy*, *sadness* and *surprise*; Plutchik (1980) shared the same set with two additions: *anticipation* and *trust*. Instead of relying on discrete categories, Russell (1980) proposed the circumplex model where every emotion can be described by three continuous values: *arousal*, *dominance* and *valence*. More recent studies in psychology use more fine-grained sets of emotions, ranging from 12 (Cowen et al., 2019b) to 28 categories (Cowen and Keltner, 2020), devised depending on the context of the study, e.g., speech prosody and facial expressions.

However, the analysis of Demszky et al. (2020) over a fine-grained set of 28 emotions suggests that a large number of categories results in more frequent disagreements on similar classes (such as *anger* and *annoyance*, or *excitement* and *joy*) which, in turn, can lead to low inter-annotator agreement and unbalanced distributions among some of these categories. Therefore, we adopt Plutchik's *Wheel of Emotions* (Plutchik, 2001), which provides clearly distinct and well-defined coarse-grained categories, whose composition can be used to virtually describe all other fine-grained sets. Moreover, some datasets in SRL4E (Mohammad et al., 2014; Kim and Klinger, 2018; Bostan et al., 2020) already use Plutchik's or Plutchik-

based categories.

Emotions and SRL. Over the past few years, automatic systems for SRL have achieved impressive performance in identifying and labeling predicate-argument relations (Shi and Lin, 2019; Conia and Navigli, 2020; Blloshmi et al., 2021; Conia et al., 2021), and have long become useful tools in several downstream tasks, from Question Answering (He et al., 2015) to Machine Translation (Marcheggiani et al., 2018). Defined by Màrquez et al. (2008) as the task of answering the question "*Who did What to Whom, Where, When and How?*", SRL is almost a natural choice for the extraction of the semantic constituents of those events that elicit emotional states. Indeed, emotional CUES can be seen as particular types of predicates, and their semantic constituents as their arguments.

Among the currently available datasets for emotion detection, there are some that also provide this kind of more granular semantic information. In particular Aman and Szpakowicz (2007) and Liew et al. (2016) released corpora that indicate multiple emotions and their corresponding emotion CUES in each sentence; Ghazi et al. (2015) and Gao et al. (2017) indicate the cause of an emotion, with the latter providing such annotations both in English and in Chinese. Finally, Mohammad et al. (2014), Mohammad et al. (2015), Kim and Klinger (2018) and Bostan et al. (2020) provide annotations for emotion CUES, EXPERIENCERS, TARGETS and STIMULI, employing, however, different sets of emotions in different domains. This means that the results of a system trained on one of these datasets cannot be compared against the results of another system trained on a different dataset, emphasizing the need for a unified framework to train and evaluate future approaches to this task. This is also evidenced by the success of existing works, e.g. Bostan and Klinger (2018) for sentence-level Emotion Classification and Raganato et al. (2017) for Word Sense Disambiguation. In SRL4E, not only do we aggregate the resources under the same task formulation, but we also manually correct their inconsistencies and unify the different emotion schemes.

3 SRL4E

In this Section, we introduce SRL4E. We first describe the categories of emotions and the format of the semantic roles we adopt to unify the annotation scheme of the original datasets. Next, we provide a short overview of the datasets included in SRL4E. Finally, we give a formal definition of the task.

3.1 Cue, Experiencer, Target, Stimulus

The task of SRL (Gildea and Jurafsky, 2000) is aimed at identifying, given an input sentence, who or what the participants are in an action or event denoted by a predicate. As mentioned in Section 2, this is comparable to answering the question "*Who* did *What* to *Whom*, *Where*, *When* and *How*?" (Màrquez et al., 2008). When it comes to emotions, however, the task does not necessarily revolve around an action, but more precisely around an *emotional cue*, a word or an expression that acts as a trigger for an emotion. Therefore, it would be more appropriate to reformulate the question as: "*Who* feels *What*, *towards Whom* and *Why*?". To answer this question, we first need to define a set of semantic roles, i.e., semantic relations that can exist between an emotion CUE and its semantic constituents. Following previous work (Mohammad et al., 2014; Bostan et al., 2020), we take a subset of semantic roles, namely, EXPERIENCER, TARGET and STIMULUS, from those defined in the "Emotion" semantic frame of FrameNet (Baker et al., 1998). While the use of thematic roles allows for human-readable labels (Kipper Schuler, 2005; Di Fabio et al., 2019), we also provide their respective definitions in Table 1.

3.2 Choosing a common set of emotions

In psychology, the debate on which categories are best suited for describing emotions is still open (Barrett et al., 2018; Cowen and Keltner, 2018; Cowen et al., 2019a). There are numerous studies that try to tackle this problem, and some of the most authoritative were briefly described in Section 2, above. In this work, we adopt Plutchik's Wheel of Emotions (Plutchik, 1980, 2001) to standardize the heterogeneous emotion categories used in the various datasets. Plutchik's Wheel of Emotions is composed of a coarse-grained set of 8 basic emotions: *anger*, *fear*, *sadness*, *disgust*, *surprise*, *anticipation*, *trust*, and *joy*. These emotions can be compounded into "dyads" which express the much wider range of human feelings, with the advantage of maintaining a solid and unambiguous base set. For example, combining *anticipation* together with *joy* describes the emotion of *optimism*, whereas *anticipation* with *sadness* describes *pessimism*. Further compositions are described in Appendix A.

SRL4E includes 6 datasets:

Role	Definition
CUE	Trigger word or expression that describes (even implicitly) an emotion.
EXPERIENCER	Person or entity that feels or experiences the emotion identified by the CUE.
TARGET	Person or entity towards whom/which the emotion identified by the CUE is directed.
STIMULUS	Entity, action or event that causes the emotion identified by the CUE.
Emotions	*anger*, *anticipation*, *disgust*, *fear*, *joy*, *sadness*, *surprise*, *trust*, and *other*.

Table 1: Definitions of semantic roles (CUE, EXPERIENCER, TARGET, STIMULUS) and emotion categories we use in SRL4E.

- Kim and Klinger (2018) and Bostan et al. (2020) use Plutchik's or Plutchik-based emotions;
- Aman and Szpakowicz (2007) and Gao et al. (2017) use Ekman's or Ekman-based emotions, which are a subset of Plutchik's set and can be directly mapped to it;
- Mohammad et al. (2014) use 19 emotions, but provide a mapping to Plutchik's emotions;
- Liew et al. (2016) use 28 emotions for which we provide a mapping to Plutchik's emotions.

We provide a more detailed description of each dataset in Section 3.3.

As a further contribution, we produce an alignment of each set of emotions to a sentiment polarity – positive, negative, neutral, or other (used when polarity cannot be inferred based on the emotion category) – to allow SRL4E also to be used to train and evaluate a system on Semantic Role Labeling for Sentiments.

Resource	Original	SRL4E	%
Blogs	5,202	4,855	93.3
Elections	1,385	1,024	73.9
EmoTweet	15,553	15,553	100.0
GNE	5,000	5,000	100.0
NTCIR (ZH)	2,022	1,956	96.7
NTCIR (EN)	1,826	1,796	98.4
REMAN	1,720	1,705	99.1
All	32,708	31,889	97.5

Table 2: Original/new sizes after conversion to SRL4E.

Resource	cue	stim.	exp.	targ.
Blogs	✔	–	–	–
Elections	✔	✔	✔	✔
EmoTweet	✔	–	–	–
GNE	✔	✔	✔	✔
NTCIR	✔	✔	–	–
REMAN	✔	✔	✔	✔

Table 3: Annotations for each of the datasets making up SRL4E.

3.3 Sources

In the following, we describe the datasets that we included in SRL4E. For each dataset, we provide general information, including source, domain, format and tagging scheme. We also indicate where we intervened manually to identify and correct errors such as typos, format errors and inconsistencies. Table 2 reports the sizes of the original and converted datasets in SRL4E. Table 3 summarizes which annotations form part of the original corpora and, therefore, which ones are also part of SRL4E. We report the license, availability and link of each resource in Appendix B.

Blogs. This dataset, proposed by Aman and Szpakowicz (2007), consists of 5,202 sentences, extracted from 173 online blog posts. Each sentence is annotated using Ekman's six emotion categories and *no emotion*, along with intensities. The words or spans that indicate emotions are marked, allowing us to remap them to the CUE in our unified format. The dataset was annotated by two experts. For each sample, we decided to consider only the CUES that were annotated with the same emotion by both annotators. Where possible, we manually identified and corrected some annotations containing typos.

Elections. This dataset, introduced by Mohammad et al. (2014, 2015), includes 1,385 unique tweets related to the 2012 US presidential election and collected using the Twitter API. The tweets were annotated via crowdsourcing using an informative tagging scheme which comprised not only

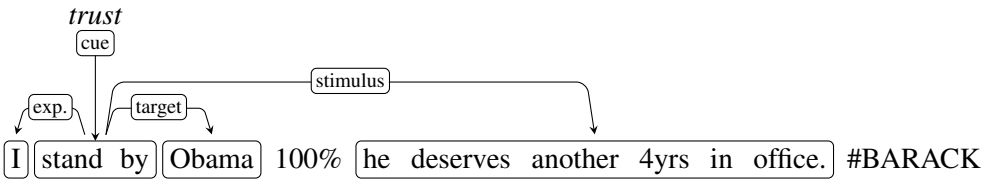

Figure 1: A sentence from the Elections dataset (Mohammad et al., 2014) using the SRL4E format. Here, the CUE expression is "stand by", and its associated emotion is Trust. The participants to the emotion are "I" (EXPERIENCER of Trust), "Obama" (TARGET of Trust), and "he deserves another 4yrs in office" (STIMULUS of Trust).

a set of 19 emotions, but also other features such as emotion intensity, valence, purpose, style, CUE, EXPERIENCER, TARGET and STIMULUS. Each sample was annotated by multiple people, i.e., each sample appears more than once with different annotations, one for each annotator. We adjudicated role spans by majority voting, discarding all the tweets with conflicting annotations.

EmoTweet. EmoTweet, presented by Liew et al. (2016), is the largest dataset that we include in our unified resource. It comprises 15,553 tweets, collected through the Twitter API using various sampling strategies (e.g., by user, by topic, random, etc.) and annotated via crowdsourcing. The original tagging scheme of this dataset features 28 emotion categories along with valence and arousal. For each emotion, the CUES are indicated and are easily mappable to our unified format. However, a mapping to Plutchik's emotions is not provided by the authors, so we formulated a conversion scheme based on the similarity of the emotion categories with those from other works that are instead mapped to Plutchik's emotions, such as Demszky et al. (2020). In addition, we also intervened to identify and manually correct some typos in the annotations.

GNE. GoodNewsEveryone, proposed by Bostan et al. (2020), is a dataset composed of 5,000 news headlines from 82 sources, annotated via crowdsourcing. It is labeled with writer and reader emotions using a set of emotions derived from Pluthick's classes and is, therefore, easily mappable to the standard Plutchik set. To keep the annotations consistent with those of the other datasets in our unified framework, we considered only the writer's emotions. GNE provides annotations for every semantic role we include in our framework, namely, CUE, EXPERIENCER, TARGET and STIMULUS, making this resource highly valuable for our purposes. Whenever possible, we identified and manually corrected the annotations that contained typos.

NTCIR 13 ECA. This dataset was proposed as a part of the NTCIR 13 Emotion Cause Analysis task. It consists of 1,826 unique sentences from English novels and 2,022 unique sentences from Chinese news, annotated using Ekman's classes. Moreover, emotion keywords and causes are annotated, making them suitable to be considered, respectively, as CUE and STIMULUS in our unified format.

REMAN. Relational EMotion ANnotation, introduced by Kim and Klinger (2018), is a corpus consisting of 1,720 fictional text excerpts from Project Gutenberg. These documents were annotated using an informative tagging scheme, which included emotion categories based on Plutchik's set, CUE, EXPERIENCER, TARGET, STIMULUS, named entities, events and coreferences, making it another desirable dataset for our unified framework. For some sentences, we automatically identified and manually corrected some typos in order to increase the overall quality of this dataset.

3.4 Task Definition

Here we provide a more formal definition of the SRL4E task. Unlike the majority of previous work on emotion detection, instead of assigning an emotion to a sentence, we associate each emotion with a CUE. In this way, in each sentence, more than one CUE can be identified and associated with its corresponding emotion category and semantic roles, allowing the coexistence of multiple emotions, EXPERIENCERS, TARGETS and STIMULI in the same sentence. A visual representation of the relationship between CUE, emotion category and roles is shown in Figure 1. To the best of our knowledge, other than SRL4E, Liew et al. (2016) and Kim and Klinger (2018) are the only approaches that leverage CUES to model the presence of multiple emotions in a sentence.

In general, the task of Semantic Role Labeling

Resource	text	cue	exp.	targ.	stim.
Blogs	13.95	2.09	–	–	–
Elections	16.66	8.29	0.02	2.19	7.92
EmoTweet	15.94	3.72	–	–	–
GNE	11.32	1.44	1.80	4.64	7.19
NTCIR (EN)	57.71	1.37	–	–	8.72
REMAN	59.15	1.84	1.53	3.81	7.36
All	19.87	2.86	1.46	4.18	7.55

Table 4: Average length (in words) of text and roles of the corpora in SRL4E (only English ones are reported). Note: EXPERIENCER average length is less than one because it is very often labeled as "author" and in this case it does not appear explicitly in the text.

for Emotions can be divided into three key steps: CUE identification, emotion classification and role identification. While there are no hard constraints on the order of these steps, we believe that CUE identification should be done first since its output will serve as the input of the other two steps, however, we also believe that our framework could be a step towards the development of joint approaches that solve the three steps at the same time.

Cue identification. As we described earlier, the CUE acts similarly to a predicate in SRL. Indeed, the main objective of CUE identification is to recognize where and how many emotions are present in a sentence, and what their trigger words or expressions are. The output of this step consists of a set of CUES, each corresponding to an emotion in the text, as illustrated in Figure 1.

Emotion classification. Traditional approaches in emotion classification take as input a sentence and output the emotion class corresponding to that sentence. In SRL4E, instead, given a pair (sentence, CUE) we want to classify the emotion expressed in the sentence by the indicated input CUE. Note that the result of this approach is not necessarily the same as a sentence-level approach.

Role identification. As previously stated, SRL aims at identifying the semantic constituents of an action expressed by a predicate. In SRL4E, instead, we are interested in identifying the actants of an emotional event which is hinted at by the CUE. Therefore a CUE can be considered in the same way as a predicate in SRL and role identification consists in identifying all those spans of text that have a semantic relationship – EXPERIENCER, TARGET, STIMULUS – with the CUE.

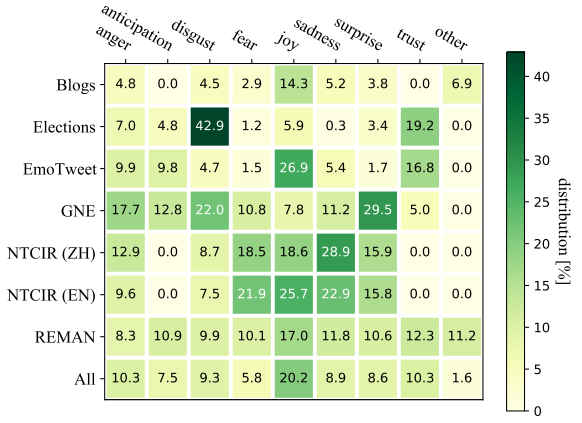

Figure 2: Distribution of categories in the corpora of SRL4E. Note: for each dataset the sum is not necessarily 100%, since there are samples where more than one CUE appears, and others where no CUES (and therefore no emotions) appear at all.

4 Data Analysis

Emotion classes distribution. Depending on the dataset, the distribution of emotion classes changes drastically, as illustrated in Figure 2. For example, in Elections, which contains random tweets related to an American election campaign, almost 45% of samples are tagged with *disgust*, as one might expect: this is because many of the tweets in question tend to discredit the opposing party; similarly, the second most used class is *trust* in the tweets in favor of candidates. Another interesting example is GNE, where the most frequent category is *surprise*, highlighting the sensationalistic tone typically found in newspaper headlines. It is worth noting that, in contrast to each individual dataset, our unified dataset includes a fairly balanced distribution between categories, where the only category that appears more often is *joy* (20%), while all the others are between 6% and 10%, approximately.

Other statistics. The statistics reported in Table 4 show the heterogeneity of the resources included in our framework, with very different text and role lengths. In fact, datasets containing sentences from similar domains share similar values. For example, REMAN and the English version of NTCIR both come from novels and they have comparable text lengths, from 58 to 59 words on average. Instead, Blogs (from online blog posts), Elections and EmoTweet (from tweets) have much shorter sentences, from 14 to 16 words, approximately.

Table 4 also shows that almost all CUEs are very short, usually around 1-2 words; only those datasets involving tweets have a much higher value. In fact, in EmoTweet and Elections, CUEs contain 4 and 8 words on average, respectively, due to their dense emotional content and, therefore, their larger number of trigger expressions. It is interesting to note that all datasets feature a similar average length for STIMULI, regardless of the domain.

Borderline examples. SRL4E's formulation is based on the presence of CUEs within sentences, which are seen as the trigger of the emotion in that context. This formulation particularly suits those domains where emotions are expressed explicitly, such as GNE, NTCIR and REMAN. However, handling CUEs becomes non-trivial in some situations, for example in social networks (Elections and EmoTweet) and blog posts (Blogs). In these contexts, language features numerous implicit references and ironic content, where the mere presence of an emoji or a particular punctuation mark completely changes the context. In our task formulation, the presence of a CUE is a fundamental requirement even if it may be difficult to identify, as we want to be able to model multiple, sometimes opposite, emotions in the same sentence. Here is an example:

- "@user Quieter. My sis, brother in law and habibti are going back to Ireland this afternoon [:/ CUE] Tennis doubles [sounds fun CUE]! [Enjoy CUE]! #Juice!"

In this case, the *sadness* emotion is associated only with the first CUE, which is ";/", while the *joy* emotion is associated with the other two. Even if a CUE is composed only of punctuation marks (or emojis), it may still be the only useful signal for disambiguating the emotion, or for separating the presence of multiple emotions in the same sentence.

5 Experiments

In this Section, we analyze the benefits that our unified framework can bring to a neural model, based on recent contextualized representations from a pretrained language model.

5.1 Emotion Classification

The main roadblock to the development of neural models for Semantic Role Labeling for Emotions is the heterogeneity of the emotion labels employed by each currently available dataset. Therefore, we first evaluate the benefits that a unified framework brings in emotion classification. Note that, differently from traditional sentence-level Emotion Detection, here we are interested in assigning an emotion to a given (sentence, CUE) pair, so as to allow a sentence to be assigned different emotions depending on the CUE considered.

Model description. We design a simple neural baseline composed mainly of a BERT-based word representation module and a stack of BiLSTM layers. Given an input sentence $\mathbf{w}$ and a pre-identified CUE $\mathbf{c}$, the two are concatenated as an input sequence $\mathbf{s}$ = [CLS] $\mathbf{w}$ [SEP] $\mathbf{c}$ [SEP] and fed into the BERT-based word representation module, obtaining a sequence of word encodings $\mathbf{e} = \text{BERT}(\mathbf{s})$. These word encodings are further processed by a stack of 2 BiLSTM layers with hidden size 512 to obtain a new sequence of output encodings $\mathbf{o} = \text{BiLSTM}(\mathbf{e})$. Finally, the output encoding $\mathbf{o}_{\text{[CLS]}}$ corresponding to the [CLS] token is fed into a linear classifier which outputs the emotions corresponding to the (sentence, CUE) pair.

Each model configuration is trained to minimize a binary cross-entropy loss for emotion classification (more than one emotion can be assigned to a given input), for a total of 20 epochs with Adam and a learning rate of 10^{-3}, leaving the weights of the underlying language model frozen.

Results. Table 5 shows the results of our system on emotion classification. First, our unified framework reveals that a system trained on a single dataset can achieve good results on the test set of the same dataset, i.e., on an in-domain evaluation, but is not able to perform as well on other datasets, i.e., on out-of-domain evaluations. Instead, the same system trained jointly on the datasets of SRL4E is able not only to perform consistently across all the test sets, but also to improve over the same system trained on in-domain data only, demonstrating empirically the effectiveness of employing a unified scheme for emotion classification. This is not a given, since each dataset differs – sometimes significantly – from the others in domain and linguistic register. On average, when using multilingual-BERT as the underlying language model, our unified framework provides an improvement of 11.2% in F1 score over EmoTweet, the second best dataset (64.3% against 53.1%). Moreover, Table 5 shows that our unified framework

Model	Trained on							Evaluated on (F1 score)							
	BL	EL	ET	GN	N/E	N/Z	RE	BL	EL	ET	GN	N/E	N/Z	RE	ALL
multilingual-BERT	✔	–	–	–	–	–	–	51.0	13.3	38.6	15.3	24.6	11.1	21.8	29.9
	–	✔	–	–	–	–	–	9.2	40.5	21.7	15.6	8.7	7.9	13.7	17.2
	–	–	✔	–	–	–	–	49.9	32.2	**76.7**	20.1	48.8	22.8	38.2	53.1
	–	–	–	✔	–	–	–	34.3	25.5	30.3	29.0	29.0	18.3	23.3	28.6
	–	–	–	–	✔	–	–	42.1	10.8	34.2	4.0	30.2	11.9	20.1	26.0
	–	–	–	–	–	✔	–	8.9	5.7	17.5	2.6	21.4	22.8	9.2	13.9
	–	–	–	–	–	–	✔	35.4	7.8	22.1	4.8	16.1	2.8	23.5	17.8
	✔	✔	✔	✔	✔	✔	✔	**65.9**	**40.7**	74.6	**33.7**	**78.5**	**77.8**	**54.1**	**64.3**

Table 5: F1 scores on **emotion classification**. Training a model on the union of all the datasets brings consistent – sometimes very large – improvements, especially on bilingual emotion classification. **BL:** Blogs. **EL:** Elections. **ET:** EmoTweet. **GN:** GNE. **N/E**: NTCIR in English. **N/Z**: NTCIR in Chinese. **RE:** REMAN.

Model	Trained on							Evaluated on (F1 score)							
	BL	EL	ET	GN	N/E	N/Z	RE	BL	EL	ET	GN	N/E	N/Z	RE	ALL
multilingual-BERT	✔	–	–	–	–	–	–	**58.8**	22.4	32.6	29.3	39.8	31.5	41.4	34.2
	–	✔	–	–	–	–	–	19.2	**50.5**	33.5	21.7	9.1	11.5	16.5	23.1
	–	–	✔	–	–	–	–	40.3	32.3	**60.7**	18.3	28.0	34.5	38.7	47.3
	–	–	–	✔	–	–	–	37.9	16.1	19.8	**57.5**	45.3	6.1	36.7	26.5
	–	–	–	–	✔	–	–	19.3	1.6	6.1	8.6	53.7	10.9	20.3	10.2
	–	–	–	–	–	✔	–	4.0	1.1	1.7	0.0	12.8	55.3	5.7	9.8
	–	–	–	–	–	–	✔	38.0	18.7	24.6	27.0	42.5	29.7	**50.8**	28.3
	✔	✔	✔	✔	✔	✔	✔	50.8	42.3	58.9	55.2	**59.2**	**61.6**	49.0	**56.5**

Table 6: F1 scores of our baseline model on CUE **identification**. **BL:** Blogs. **EL:** Elections. **ET:** EmoTweet. **GN:** GNE. **N/E**: NTCIR in English. **N/Z**: NTCIR in Chinese. **RE:** REMAN.

allows our system to improve in bilingual emotion classification (77.8% against 22.8% in F1 score on ALL).

5.2 Cue Identification

We now turn to CUE identification, where we aim to find every CUE in an input sentence. We frame this subtask as a BIO-tagging problem and devise a neural model to highlight the benefits of our unified framework in this task.

Model description. For CUE identification, we use a similar system architecture to the one we used for emotion classification. However, this time the input of the BERT-based word representation module is just the input sentence, whereas the output is a sequence of BIO tags. Specifically, the output encodings $\mathbf{o} = \mathbf{o}_1, \ldots, \mathbf{o}_n$ produced by the last BiLSTM layer are given to a classifier which learns to predict B-cue, I-cue or O.

Results. As one can see in Table 6, similarly to what we observed in emotion classification, our unified framework highlights how a model trained on a single dataset is not robust to out-of-domain evaluations. Instead, the same model trained on all the datasets in SRL4E shows consistent results across all the test sets, providing a significant improvement in F1 score over the second best dataset, EmoTweet (56.5% against 47.3% in F1 score on ALL, with an absolute improvement of 9.2%).

5.3 Role Identification

Model description. For role identification, we use an approach similar to that for CUE identification. Indeed, similarly to CUE identification, we model role identification as a BIO-tagging problem, with the only difference being that we provide the pre-identified CUE in input, i.e., the input sequence is $\mathbf{s}$ = [CLS] $\mathbf{w}$ [SEP] $\mathbf{c}$ [SEP], where $\mathbf{w}$ is the input sentence and $\mathbf{c}$ is the CUE span.

Results. We find that our results on the identification of each role are in line with the results from CUE identification, leading us to draw similar conclusions (see Tables 7, 8 and 9). In general, we see a familiar pattern in which training our baseline model on a single dataset results in good perfor-

Model	Trained on					Evaluated on (F1 score)					
	EL	GN	N/E	N/Z	RE	EL	GN	N/E	N/Z	RE	ALL
multilingual-BERT	✔	–	–	–	–	52.8	62.7	24.5	25.9	14.1	32.2
	–	✔	–	–	–	42.4	75.8	21.9	22.4	13.8	37.1
	–	–	✔	–	–	31.6	40.8	50.4	12.9	20.1	30.3
	–	–	–	✔	–	16.5	16.5	20.1	56.2	15.9	38.5
	–	–	–	–	✔	9.8	9.0	24.4	3.8	**26.4**	10.3
	✔	✔	✔	✔	✔	**54.5**	**76.3**	**52.7**	**57.8**	16.6	**62.5**

Table 7: F1 scores of our baseline model on STIMULUS **identification**. **BL:** Blogs. **EL:** Elections. **ET:** EmoTweet. **GN:** GNE. **N/E**: NTCIR in English. **N/Z**: NTCIR in Chinese. **RE:** REMAN.

	Trained on			Evaluated on (F1 score)			
	EL	GN	RE	EL	GN	RE	ALL
m-BERT	✔	–	–	**98.27**	0.18	0.00	10.23
	–	✔	–	2.13	71.15	28.07	56.77
	–	–	✔	3.01	43.45	55.72	43.19
	✔	✔	✔	**98.27**	**71.86**	**58.31**	**71.54**

Table 8: F1 scores on EXPERIENCER **identification**. **EL:** Elections. **GN:** GNE. **RE:** REMAN.

	Trained on			Evaluated on (F1 score)			
	EL	GN	RE	EL	GN	RE	ALL
m-BERT	✔	–	–	**63.33**	13.03	11.94	17.84
	–	✔	–	14.31	**55.45**	14.21	42.52
	–	–	✔	32.09	29.77	40.79	31.52
	✔	✔	✔	54.44	50.13	**43.14**	**49.58**

Table 9: F1 scores on TARGET **identification**. **EL:** Elections. **GN:** GNE. **RE:** REMAN.

mances on that specific dataset, but with significantly lower results on out-of-domain data.

5.4 Result Analysis

Results generally benefit from a unified resource. For instance, emotion classification and STIMULUS identification almost always struggle in out-of-domain evaluations, while they perform better when the model is trained on all the datasets at the same time.

The only exception is CUE identification: when our model is trained on all the data in SRL4E, the performance drops when measured on each dataset separately. This is to be expected: while STIMULI follow a similar syntactic pattern across domains, CUES appear in very different forms (e.g., Twitter usually contains highly informal language with explicit emotions, while news headlines tend to try to describe events objectively, making emotions more implicit). Instead, when the datasets share a similar domain, the model is able to generalize well, even in cross-lingual settings (such as the English and Chinese versions of NTCIR), highlighting once again the advantages of our unified framework.

6 Conclusion and Future Work

Recently, the study of emotions in NLP has been gaining interest, due to their potential not only for application to downstream tasks, but also for enhancing the interpretability of automatic outputs, especially when emotions are accompanied by information about their semantic constituents, i.e., their experiencers, targets and stimuli. However, recent efforts to provide manually annotated data for emotions and their semantic constituents have been heterogeneous in their annotation scheme, making it difficult to train, evaluate, and compare novel approaches.

In this paper, we aimed at addressing these issues and presented a unified framework for the Semantic Role Labeling of Emotions (SRL4E). Our framework collects, cleans, and unifies the annotation schemes of six datasets that provide information about emotions and their semantic roles, making it easy to train and evaluate existing and future systems. We conducted several experiments to demonstrate empirically that our unified scheme is beneficial in each subtask, namely, emotion classification and role (experiencer, target, stimulus) identification, especially in bilingual settings (English-Chinese). With SRL4E, we hope to stimulate future research in this complex area at the intersection of Emotion Detection and Semantic Role Labeling. We release the software to reproduce the benchmark and our experiments at https://github.com/SapienzaNLP/srl4e.

Acknowledgments

The authors gratefully acknowledge the support of the ERC Consolidator Grant MOUSSE No. 726487 and the European Language Grid project No. 825627 (Universal Semantic Annotator, USeA) under the European Union's Horizon 2020 research and innovation programme.

This work was supported in part by the MIUR under grant "Dipartimenti di Eccellenza 2018-2022" of the Department of Computer Science of Sapienza University of Rome.

References

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from Text: Machine Learning for Text-based Emotion Prediction. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 579–586, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Saima Aman and Stan Szpakowicz. 2007. Identifying Expressions of Emotion in Text. In *Text, Speech and Dialogue*, Lecture Notes in Computer Science, pages 196–205, Berlin, Heidelberg. Springer.

Saima Aman and Stan Szpakowicz. 2008. Using Roget's Thesaurus for Fine-grained Emotion Recognition. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet Project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, Montreal, Quebec, Canada. Association for Computational Linguistics.

Lisa Feldman Barrett, Zulqarnain Khan, Jennifer Dy, and Dana Brooks. 2018. Nature of Emotion Categories: Comment on Cowen and Keltner. *Trends in Cognitive Sciences*, 22(2):97–99.

Rexhina Blloshmi, Simone Conia, Rocco Tripodi, and Roberto Navigli. 2021. Generating senses and roles: An end-to-end model for dependency- and span-based semantic role labeling. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3786–3793.

Laura Ana Maria Bostan, Evgeny Kim, and Roman Klinger. 2020. GoodNewsEveryone: A Corpus of News Headlines Annotated with Emotions, Semantic Roles, and Reader Perception. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1554–1566, Marseille, France. European Language Resources Association.

Laura-Ana-Maria Bostan 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, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Sven Buechel and Udo Hahn. 2017. EmoBank: Studying the Impact of Annotation Perspective and Representation Format on Dimensional Emotion Analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585, Valencia, Spain. Association for Computational Linguistics.

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.

Dushyant Singh Chauhan, Dhanush S R, Asif Ekbal, and Pushpak Bhattacharyya. 2020. Sentiment and emotion help sarcasm? a multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4351–4360, Online. Association for Computational Linguistics.

Simone Conia, Andrea Bacciu, and Roberto Navigli. 2021. Unifying cross-lingual semantic role labeling with heterogeneous linguistic resources. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 338–351, Online.

Simone Conia and Roberto Navigli. 2020. Bridging the gap in multilingual Semantic Role Labeling: a language-agnostic approach. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1396–1410, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Alan Cowen, Disa Sauter, Jessica L. Tracy, and Dacher Keltner. 2019a. Mapping the Passions: Toward a High-Dimensional Taxonomy of Emotional Experience and Expression. *Psychological Science in the Public Interest*, 20(1):69–90. Publisher: SAGE Publications Inc.

Alan S. Cowen and Dacher Keltner. 2018. Clarifying the Conceptualization, Dimensionality, and Structure of Emotion: Response to Barrett and Colleagues. *Trends in Cognitive Sciences*, 22(4):274–276.

Alan S. Cowen and Dacher Keltner. 2020. What the face displays: Mapping 28 emotions conveyed by naturalistic expression. *The American Psychologist*, 75(3):349–364.

Alan S. Cowen, Petri Laukka, Hillary Anger Elfenbein, Runjing Liu, and Dacher Keltner. 2019b. The primacy of categories in the recognition of 12 emotions in speech prosody across two cultures. *Nature Human Behaviour*, 3(4):369–382.

CrowdFlower. 2016. Sentiment Analysis in Text - dataset by crowdflower.

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, Online. Association for Computational Linguistics.

Andrea Di Fabio, Simone Conia, and Roberto Navigli. 2019. VerbAtlas: a novel large-scale verbal semantic resource and its application to semantic role labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 627–637, Hong Kong, China.

P. Ekman. 1992. Are there basic emotions? *Psychological Review*, 99(3):550–553.

Qinghong Gao, Jiannan Hu, Ruifeng Xu, Gui Lin, Yulan He, Qin Lu, and Kam-Fai Wong. 2017. Overview of NTCIR-13 ECA Task. page 6.

Diman Ghazi, Diana Inkpen, and Stan Szpakowicz. 2015. Detecting Emotion Stimuli in Emotion-Bearing Sentences. In *Computational Linguistics and Intelligent Text Processing*, volume 9042, pages 152–165. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science.

Daniel Gildea and Daniel Jurafsky. 2000. Automatic Labeling of Semantic Roles. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 512–520, Hong Kong. Association for Computational Linguistics.

Luheng He, Mike Lewis, and Luke Zettlemoyer. 2015. Question-answer driven semantic role labeling: Using natural language to annotate natural language. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 643–653, Lisbon, Portugal.

Pere-Lluís Huguet Cabot, David Abadi, Agneta Fischer, and Ekaterina Shutova. 2021. Us vs. Them: A Dataset of Populist Attitudes, News Bias and Emotions. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1921–1945, Online. Association for Computational Linguistics.

Evgeny Kim and Roman Klinger. 2018. Who Feels What and Why? Annotation of a Literature Corpus with Semantic Roles of Emotions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1345–1359, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Karin Kipper Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. University of Pennsylvania.

Jingsheng Lei, Yanghui Rao, Qing Li, Xiaojun Quan, and Liu Wenyin. 2014. Towards building a social emotion detection system for online news. *Future Generation Computer Systems*, 37:438–448.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Jasy Suet Yan Liew, Howard R. Turtle, and Elizabeth D. Liddy. 2016. EmoTweet-28: A Fine-Grained Emotion Corpus for Sentiment Analysis. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1149–1156, Portorož, Slovenia. European Language Resources Association (ELRA).

Vicki Liu, Carmen Banea, and Rada Mihalcea. 2017. Grounded emotions. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 477–483. ISSN: 2156-8111.

Diego Marcheggiani, Jasmijn Bastings, and Ivan Titov. 2018. Exploiting semantics in neural machine translation with graph convolutional networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 486–492, New Orleans, Louisiana. Association for Computational Linguistics.

Ilia Markov, Nikola Ljubešić, Darja Fišer, and Walter Daelemans. 2021. Exploring stylometric and emotion-based features for multilingual cross-domain hate speech detection. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 149–159, Online.

Saif Mohammad. 2012. #Emotional Tweets. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 246–255, Montréal, Canada. Association for Computational Linguistics.

Saif Mohammad and Felipe Bravo-Marquez. 2017. Emotion Intensities in Tweets. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 65–77, Vancouver, Canada. Association for Computational Linguistics.

Saif Mohammad, Xiaodan Zhu, and Joel Martin. 2014. Semantic Role Labeling of Emotions in Tweets. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 32–41, Baltimore, Maryland. Association for Computational Linguistics.

Saif M. Mohammad, Xiaodan Zhu, Svetlana Kiritchenko, and Joel Martin. 2015. Sentiment, emotion, purpose, and style in electoral tweets. *Information Processing & Management*, 51(4):480–499.

Lluís Màrquez, Xavier Carreras, Kenneth C. Litkowski, and Suzanne Stevenson. 2008. Special Issue Introduction: Semantic Role Labeling: An Introduction to the Special Issue. *Computational Linguistics*, 34(2):145–159.

Laura Ana Maria Oberländer, Kevin Reich, and Roman Klinger. 2020. Experiencers, Stimuli, or Targets: Which Semantic Roles Enable Machine Learning to Infer the Emotions? In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 119–128, Barcelona, Spain (Online). Association for Computational Linguistics.

ROBERT Plutchik. 1980. Chapter 1 - A GENERAL PSYCHOEVOLUTIONARY THEORY OF EMOTION. In Robert Plutchik and Henry Kellerman, editors, *Theories of Emotion*, pages 3–33. Academic Press.

Robert Plutchik. 2001. The Nature of Emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*, 89(4):344–350. Publisher: Sigma Xi, The Scientific Research Society.

Daniel Preoţiuc-Pietro, H. Andrew Schwartz, Gregory Park, Johannes Eichstaedt, Margaret Kern, Lyle Ungar, and Elisabeth Shulman. 2016. Modelling Valence and Arousal in Facebook posts. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 9–15, San Diego, California. Association for Computational Linguistics.

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017. Word Sense Disambiguation: A Unified Evaluation Framework and Empirical Comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 99–110, Valencia, Spain. Association for Computational Linguistics.

Santhosh Rajamanickam, Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2020. Joint Modelling of Emotion and Abusive Language Detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4270–4279, Online. Association for Computational Linguistics.

James A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178. Place: US Publisher: American Psychological Association.

K. R. Scherer and H. G. Wallbott. 1994. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of Personality and Social Psychology*, 66(2):310–328.

Hendrik Schuff, Jeremy Barnes, Julian Mohme, Sebastian Padó, and Roman Klinger. 2017. Annotation, Modelling and Analysis of Fine-Grained Emotions on a Stance and Sentiment Detection Corpus. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 13–23, Copenhagen, Denmark. Association for Computational Linguistics.

Peng Shi and Jimmy Lin. 2019. Simple BERT models for Relation Extraction and Semantic Role Labeling. *CoRR*, abs/1904.05255.

Tiberiu Sosea and Cornelia Caragea. 2020. CancerEmo: A Dataset for Fine-Grained Emotion Detection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8892–8904, Online. Association for Computational Linguistics.

Carlo Strapparava and Rada Mihalcea. 2007. SemEval-2007 Task 14: Affective Text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74, Prague, Czech Republic. Association for Computational Linguistics.

Shabnam Tafreshi and Mona Diab. 2018. Sentence and Clause Level Emotion Annotation, Detection, and Classification in a Multi-Genre Corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Wikimedia. 2020. Plutchik dyads. [Online; accessed 17-May-2021].

Emily Öhman, Marc Pàmies, Kaisla Kajava, and Jörg Tiedemann. 2020. XED: A Multilingual Dataset for Sentiment Analysis and Emotion Detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6542–6552, Barcelona, Spain (Online). International Committee on Computational Linguistics.

A Plutchik's Wheel of Emotions

Pluthick's basic emotions – *anger*, *anticipation*, *disgust*, *fear*, *joy*, *sadness*, *surprise* and *trust* – can be compounded into "dyads" to form even more complex feelings. The compositions are shown in Figure 3. These can be used to describe the emotional context in which basic emotions are not enough, but a more fine-grained set is needed.

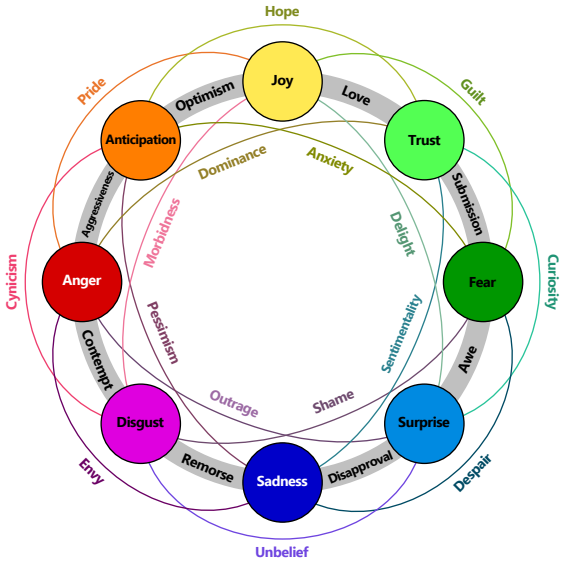

Figure 3: Pluthick's emotion dyads (Wikimedia, 2020)

.

B Sources: Additional Information

In this Section, we list the license and availability of each of the six resources included in SRL4E with a link to where to download them, if available:

- **Blogs.** The license is not specified, but it is available for research purposes upon request to the authors;
- **Elections.** The license is not specified; it is freely available online[1] and can be used for research purposes;
- **EmoTweet.** The license is not specified, but it is available for research purposes upon request to the authors;
- **GNE.** This dataset is freely available online[2] under CC BY 4.0 license;
- **NTCIR 13 ECA.** The license is not specified and the download page is no longer online, but a snapshot can be accessed using Internet Archive;[3]
- **REMAN.** This dataset is freely available online[4] under CC BY 4.0 license.

A summary of the above information is reported in table 10.

C Comparison

Oberländer et al. (2020) did, in fact, aggregate a similar set of corpora (which is now actually a subset of SRL4E) for addressing emotion classification, however, we stress that our task formulation is different: their goal is to study which semantic roles allow models to infer emotions.

Instead, SRL4E proposes a novel task formulation, together with a unified dataset for such a task. Moreover, as opposed to the above mentioned work, SRL4E assigns emotions to CUES, not to whole sentences. Finally, our dataset is larger, treats Emotion Classification as a multi-label classification task (i.e. multiple emotions can be assigned to the same CUE) and features a manual correction of annotations issues (e.g. typos, inconsistencies).

D Experiments: Additional Results

Additional experiments were conducted to compare the performance of the models in monolingual and multilingual settings. Results for emotion classification and sentiment classification are reported in Tables 11 and 12, respectively.

E SRL4E Format Example

SRL4E generates a set of JSON files. An example of how a sample is represented in SRL4E format is shown in Listing 1.

[1] http://saifmohammad.com/WebPages/SentimentEmotionLabeledData.html

[2] https://www.ims.uni-stuttgart.de/en/research/resources/corpora/goodnewseveryone/

[3] https://web.archive.org/web/20170913034355/http://hlt.hitsz.edu.cn/ECA.html

[4] https://www.ims.uni-stuttgart.de/en/research/resources/corpora/reman/

Resource	Source	Lic.	Link
Blogs	Aman and Szpakowicz (2007)	R-R	–
Elections	Mohammad et al. (2014)	D-R	Link
EmoTweet	Liew et al. (2016)	R-R	–
GNE	Bostan et al. (2020)	D-C	Link
NTCIR	Gao et al. (2017)	D-U	Link
REMAN	Kim and Klinger (2018)	D-C	Link

Table 10: License information, availability and link for each dataset that is part of SRL4E. **R-R**: available upon **R**equest for **R**esearch only purposes; **D-R**: available online for **D**ownload for **R**esearch only purposes; **D-U**: available online for **D**ownload with **U**nknown license; **D-C**: available online for **D**ownload under **C**C-BY 4.0 license.

	Trained on							Evaluated on (F1 score)							
Model	**BL**	**EL**	**ET**	**GN**	**N/E**	**N/Z**	**RE**	**BL**	**EL**	**ET**	**GN**	**N/E**	**N/Z**	**RE**	**ALL**
BERT-base	✔	–	–	–	–	–	–	57.9	24.0	40.4	20.1	25.9	–	32.4	35.0
	–	✔	–	–	–	–	–	16.7	49.4	26.8	17.8	8.1	–	17.7	22.5
	–	–	✔	–	–	–	–	52.6	36.4	**80.3**	23.6	50.5	–	41.5	58.4
	–	–	–	✔	–	–	–	35.9	37.4	30.3	**34.9**	29.6	–	29.1	32.0
	–	–	–	–	✔	–	–	49.9	14.3	39.7	17.3	71.4	–	31.2	37.4
	–	–	–	–	–	–	✔	47.2	21.7	36.6	9.5	32.5	–	38.1	31.5
	✔	✔	✔	✔	✔	–	✔	**72.6**	**56.5**	77.0	30.7	**77.3**	–	**58.4**	**65.6**
multilingual-BERT	✔	–	–	–	–	–	–	51.0	13.3	38.6	15.3	24.6	11.1	21.8	29.9
	–	✔	–	–	–	–	–	9.2	40.5	21.7	15.6	8.7	7.9	13.7	17.2
	–	–	✔	–	–	–	–	49.9	32.2	**76.7**	20.1	48.8	22.8	38.2	53.1
	–	–	–	✔	–	–	–	34.3	25.5	30.3	29.0	29.0	18.3	23.3	28.6
	–	–	–	–	✔	–	–	42.1	10.8	34.2	4.0	30.2	11.9	20.1	26.0
	–	–	–	–	–	✔	–	8.9	5.7	17.5	2.6	21.4	22.8	9.2	13.9
	–	–	–	–	–	–	✔	35.4	7.8	22.1	4.8	16.1	2.8	23.5	17.8
	✔	✔	✔	✔	✔	✔	✔	**65.9**	**40.7**	74.6	**33.7**	**78.5**	**77.8**	**54.1**	**64.3**

Table 11: Comparison between *BERT-base* and *mutilingual-BERT* in terms of F1 score on **emotion classification**. *BERT-base* obtains slightly better results, but overall both the models benefit from training on multiple datasets at the same time, even if the datasets are heterogeneous in size and domain. **BL:** Blogs. **EL:** Elections. **ET:** EmoTweet. **GN:** GNE. **N/E**: NTCIR in English. **N/Z**: NTCIR in Chinese. **RE:** REMAN.

Model	Trained on							Evaluated on (Accuracy – %)							
	BL	**EL**	**ET**	**GN**	**N/E**	**N/Z**	**RE**	**BL**	**EL**	**ET**	**GN**	**N/E**	**N/Z**	**RE**	**ALL**
BERT-base	✔	–	–	–	–	–	–	81.6	60.5	69.1	66.5	55.9	–	63.5	67.6
	–	✔	–	–	–	–	–	64.3	**80.2**	70.5	71.3	61.7	–	41.0	67.4
	–	–	✔	–	–	–	–	80.6	77.9	**91.8**	75.3	88.3	–	57.1	83.1
	–	–	–	✔	–	–	–	77.0	72.1	79.2	**79.9**	77.7	–	55.1	76.9
	–	–	–	–	✔	–	–	78.1	69.8	73.0	70.9	**92.6**	–	50.6	72.9
	–	–	–	–	–	–	✔	72.5	64.0	69.6	55.7	77.7	–	**73.7**	67.5
	✔	✔	✔	✔	✔	–	✔	**82.7**	**80.2**	91.1	79.3	92.0	–	68.6	**85.3**
multilingual-BERT	✔	–	–	–	–	–	–	81.6	57.0	69.4	65.1	60.6	39.6	56.4	64.5
	–	✔	–	–	–	–	–	58.7	73.3	65.1	70.3	61.2	66.8	38.5	63.9
	–	–	✔	–	–	–	–	80.1	74.4	89.1	73.6	87.2	**76.2**	52.6	80.5
	–	–	–	✔	–	–	–	65.8	74.4	63.4	76.8	70.7	71.3	47.7	67.2
	–	–	–	–	✔	–	–	72.5	68.6	73.5	69.0	87.2	72.3	53.9	71.8
	–	–	–	–	–	✔	–	46.9	61.6	37.5	68.2	60.1	65.8	33.3	50.2
	–	–	–	–	–	–	✔	64.3	62.8	56.3	62.6	76.1	67.8	68.0	62.2
	✔	✔	✔	✔	✔	✔	✔	**84.2**	**81.4**	**90.6**	**78.0**	**92.6**	66.3	**70.5**	**83.5**

Table 12: Comparison between *BERT-base* and *mutilingual-BERT* in terms of Accuracy (%) on **sentiment classification**. Again, both the models benefit from training on multiple datasets at the same time, even if the task of sentiment classification is simpler than that of emotion classification and even if the datasets are heterogeneous in size and domain. **BL:** Blogs. **EL:** Elections. **ET:** EmoTweet. **GN:** GNE. **N/E**: NTCIR in English. **N/Z**: NTCIR in Chinese. **RE:** REMAN.

```
[
    ...
    "gne.0004953": {
        "emotions": {
            "gne.0004953.00": {
                "original_emotion": [
                    "negative_surprise"
                ],
                "plutchik_emotion": [
                    "surprise"
                ],
                "roles": {
                    "stimulus": [
                        [
                            25,
                            41
                        ]
                    ],
                    "cue": [
                        [
                            12,
                            21
                        ]
                    ],
                    "experiencer": [
                        [
                            0,
                            4
                        ]
                    ],
                    "target": [
                        [
                            25,
                            32
                        ]
                    ]
                },
                "sentiment": "negative"
            }
        },
        "text": "Barr: I Was Surprised by Mueller Decision"
    },
    ...
]
```

Listing 1: An example of an instance from the GNE dataset in the SRL4E format (JSON). Note: in this case just one CUE (with its associated emotion) is present, but multiple CUEs/Emotions may appear. A role annotation is defined by its beginning position (included) and end position (excluded) in the original text.