SpanEmo: Casting Multi-label Emotion Classification as Span-prediction

Hassan Alhuzali   **Sophia Ananiadou**

National Centre for Text Mining

Department of Computer Science, The University of Manchester, United Kingdom

{hassan.alhuzali@postgrad.,sophia.ananiadou@}manchester.ac.uk

Abstract

Emotion recognition (ER) is an important task in Natural Language Processing (NLP), due to its high impact in real-world applications from health and well-being to author profiling, consumer analysis and security. Current approaches to ER, mainly classify emotions independently without considering that emotions can co-exist. Such approaches overlook potential ambiguities, in which multiple emotions overlap. We propose a new model "SpanEmo" casting multi-label emotion classification as span-prediction, which can aid ER models to learn associations between labels and words in a sentence. Furthermore, we introduce a loss function focused on modelling multiple co-existing emotions in the input sentence. Experiments performed on the SemEval2018 multi-label emotion data over three language sets (i.e., English, Arabic and Spanish) demonstrate our method's effectiveness. Finally, we present different analyses that illustrate the benefits of our method in terms of improving the model performance and learning meaningful associations between emotion classes and words in the sentence[1].

1 Introduction

Emotion is essential to human communication, thus emotion recognition (ER) models have a host of applications from health and well-being (Alhuzali and Ananiadou, 2019; Aragón et al., 2019; Chen et al., 2018) to consumer analysis (Alaluf and Illouz, 2019; Herzig et al., 2016) and user profiling (Volkova and Bachrach, 2016; Mohammad and Kiritchenko, 2013), among others. Interest in this area has given rise to new NLP approaches aimed at emotion classification, including single-label and multi-label emotion classification. Most existing approaches for multi-label emotion classi-

---

[1]Source code is available at https://github.com/hasanhuz/SpanEmo

fication (Ying et al., 2019; Baziotis et al., 2018; Yu et al., 2018; Badaro et al., 2018; Mulki et al., 2018; Mohammad et al., 2018; Yang et al., 2018) do not effectively capture emotion-specific associations, which can be useful for prediction, as well as learning of association between emotion labels and words in a sentence. In addition, standard approaches in emotion classification treat individual emotion independently. However, emotions are not independent; a specific emotive expression can be associated with multiple emotions. The existence of association/correlation among emotions has been well-studied in psychological theories of emotions, such as Plutchik's wheels of emotion (Plutchik, 1984) that introduces the notion of mixed and contrastive emotions. For example, "joy" is close to "love" and "optimism", instead of "anger" and "sadness".

#	Sentence	GT
S1	well my day started off great the mocha machine wasn't working @ mcdonalds.	anger, disgust, joy, sadness
S2	I'm doing all this to make sure you smiling down on me bro.	joy, love, optimism

Table 1: Example Tweets from SemEval-18 Task 1. GT represents the ground truth labels.

Consider S1 in Table 1, which contains a mix of positive and negative emotions, although it is more negative oriented. This can be observed clearly via the ground truth labels assigned to this example, where the first part of this sentence only expresses a positive emotion (i.e., joy), while the other part expresses negative emotions. For example, clue words like "great" are more likely to be associated with "joy", whereas "wasn't working" are more likely to be associated with negative emotions. Learning such associations between emotion labels and

1573

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 1573–1584

April 19 - 23, 2021. ©2021 Association for Computational Linguistics

words in a sentence can help ER models to predict the correct labels. S2 further highlights that certain emotions are more likely to be associated with each other. Mohammad and Bravo-Marquez (2017) also observed that negative emotions are highly associated with each other, while less associated with positive emotions. Based on these observations, we seek to answer the following research questions: i) how to enable ER models to learn emotion-specific associations by taking into account label information and ii) how to benefit from the multiple co-existing emotions in a multi-label emotion data set with the intention of learning label correlations. Our contributions are summarised as follows:

I. a novel framework casting the task of multi-label emotion classification as a span-prediction problem. We introduce "SpanEmo" to train the model to take into consideration both the input sentence and a label set (i.e., emotion classes) for selecting a span of emotion classes in the label set as the output. The objective of SpanEmo is to predict emotion classes directly from the label set and capture associations corresponding to each emotion.

II. a loss function, modelling multiple co-existing emotions for each input sentence. We make use of the label-correlation aware loss (LCA) (Yeh et al., 2017), originally introduced by Zhang and Zhou (2006). The objective of this loss function is to maximise the distance between positive and negative labels, which is learned directly from the multi-label emotion data set.

III. a large number of experiments and analyses both at the word- and sentence-level, demonstrating the strength of SpanEmo for multi-label emotion classification across three languages (i.e. English, Arabic and Spanish).

The rest of the paper is organised as follows: section 2 describes our methodology, while section 3 discusses experimental details. We evaluate the proposed method and compare it to related methods in section 4. Section 5 reports on the analysis of results, while section 6 reviews related work. We conclude in section 7.

2 Methodology

2.1 Framework

Figure 1 presents our framework (SpanEmo). Given an input sentence and a set of classes, a base encoder was employed to learn contextualised word representations. Next, a feed forward network (FFN) was used to project the learned representa-

tions into a single score for each token. We then used the scores for the label tokens as predictions for the corresponding emotion label. The green boxes at the top of the FFN illustrate the positive label set, while the red ones illustrate the negative label set for multi-label emotion classification. We now turn to describing our framework in detail.
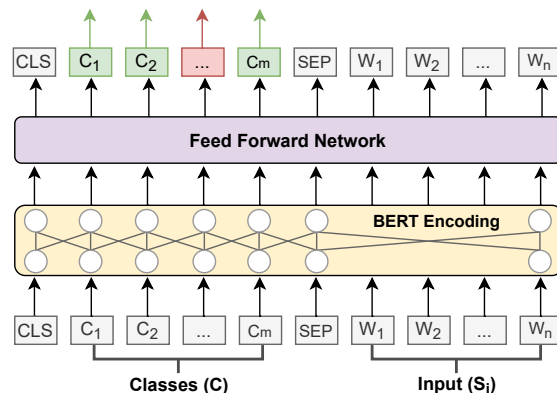

Figure 1: Illustration of our proposed framework (SpanEmo).

2.2 Our Method (SpanEmo)

Let $\{(s_i, y_i)\}_{i=1}^N$ be a set of N examples with the corresponding emotion labels of C classes, where s_i denotes the input sentence and $y_i \in \{0,1\}^m$ represents the label set for s_i. As shown in Figure 1, both the label set and the input sentence were passed into the encoder BERT (Devlin et al., 2019). The encoder received two segments: the first corresponds to the set of emotion classes, while the second refers to the input sentence. The hidden representations ($H_i \in R^{T \times D}$)[2] for each input sentence and the label set were obtained as follows:

$$H_i = \text{Encoder}([\text{CLS}] + |C| + [\text{SEP}] + \mathbf{s}_i), \quad (1)$$

where $\{[\text{CLS}], [\text{SEP}]\}$ are special tokens and $|C|$ denotes the size of emotion classes. Feeding both segments to the encoder has a few advantages. First, the encoder can interpolate between emotion classes and all words in the input sentence. Second, a hidden representation is generated both for words and emotion classes, which can be further used to understand whether the encoder can learn association between the emotion classes and words in the input sentence. Third, SpanEmo is flexible because its predictions are directly produced from the first segment corresponding to the emotion classes.

---

[2]T and D denote the input length and dimensional size, respectively.

We further introduced a feed-forward network (FFN) consisting of a non-linear hidden layer with a Tanh activation ($f_i(\mathbf{H}_i)$) as well as a position vector $p_i \in R^D$, which was used to compute a dot product between the output of f_i and p_i. As our task involved a multi-label emotion classification, we added a sigmoid activation to determine whether class$_i$ was the correct emotion label or not. It should be mentioned that the use of the position vector is quite similar to how start and end vectors are defined in transformer-based models for question-answering. Finally, the span-prediction tokens were obtained from the label segment and then compared with the ground truth labels since there was a 1-to-1 correspondence between the label tokens and the original emotion labels.

$$\hat{\mathbf{y}} = \text{sigmoid}(\text{FFN}(\mathbf{H}_i)), \qquad (2)$$

2.3 Label-Correlation Aware (LCA) Loss

Following Yeh et al. (2017), we employed the label-correlation aware loss, which takes a vector of true-binary labels (y), as well as a vector of probabilities ($\hat{y}$), as input:

$$\mathcal{L}_{\text{LCA}}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{|\mathbf{y}^0| |\mathbf{y}^1|} \sum_{(p,q) \in \mathbf{y}^0 \times \mathbf{y}^1} \exp\left(\hat{\mathbf{y}}_p - \hat{\mathbf{y}}_q\right), \quad (3)$$

where y^0 denotes the set of negative labels, while y^1 denotes the set of positive labels. $\hat{y}_p$ represents the pth element of vector $\hat{y}$. The objective of this loss function is to maximise the distance between positive and negative labels by implicitly retaining the label-dependency information. In other words, the model should be penalised when it predicts a pair of labels that should not co-exist for a given example.

2.4 Training Objective

To model label-correlation, we combined LCA loss with binary cross-entropy (BCE) and trained them jointly. This aimed to help the LCA loss to focus on maximising the distance between positive and negative label sets, while at the same time taking advantage of the BCE loss through maximising the probability of the correct labels. We experimentally observed that training our approach jointly with those two loss functions produced the best results. The overall training objective was computed as follows:

$$\mathcal{L} = (1 - \alpha)\mathcal{L}_{\text{BCE}} + \alpha \sum_{i=1}^{M} \mathcal{L}_{\text{LCA}}, \qquad (4)$$

where $\alpha \in [0, 1]$ denotes the weight used to control the contribution of each part to the overall loss.

3 Experiments

3.1 Implementation Details

We used PyTorch (Paszke et al., 2017) for implementation and ran all experiments on an Nvidia GeForce GTX 1080 with 11 GB memory. We also trained BERT$_{\text{BASE}}$ utilising the open-source Hugging-Face implementation (Wolf et al., 2019). For experiments related to Arabic, we chose "bert-base-arabic" developed by Safaya et al. (2020), while selecting "bert-base-spanish-uncased" developed by Cañete et al. (2020) for Spanish. All three models were trained on the same hyper-parameters with a fixed initialisation seed, including a feature dimension of 786, a batch size of 32, a dropout rate of 0.1, an early stop patience of 10 and 20 epochs. Adam was selected for optimisation (Kingma and Ba, 2014) with a learning rate of $2e$-5 for the BERT encoder, and a learning rate of $1e$-3 for the FFN. It should be mentioned that we tuned our method only on the validation set and further report on the analysis of the effect of parameter α in section 5.4. Table 2 summarises the hyper-parameters used in our experiments.

Parameter	Value
Feature dimension	768
Batch size	32
Dropout	0.1
Early stop patience	10
Number of epochs	20
lr-BERT	$2e$-5
lr-FFN	$1e$-3
Optimiser	Adam
Alpha (α)	0.2

Table 2: Hyper-parameter values. lr: refers to the Learning rate.

3.2 Data Set and Task Settings

In this work, we chose SemEval2018 (Mohammad et al., 2018) for our multi-label emotion classification, which is based on labelled data from tweets in English, Arabic and Spanish. The data was originally partitioned into three sets: training set (Train), validation set (Valid) and test (Test) set. Following the metrics in Mohammad et al. (2018), we run our experiments on micro F1-score, macro F1-score

and Jaccard index score[3]. Table 3 presents the summary of all three sets for each language, including the number of instances in the train, valid and test sets. In addition, the number of emotion classes and the percentage of instances with varying numbers of classes (co-existing) are included. It is worth noting that these percentages do not include the neutral instances.

Info./Lang.	English	Arabic	Spanish
Train (#)	6,838	2,278	3,561
Valid (#)	886	585	679
Test (#)	3,259	1,518	2,854
Total (#)	10,983	4,381	7,094
Classes (#)	11	11	11
1 co.emo (%)	14.36	21.38	39.11
2 co.emo (%)	40.55	39.03	42.15
3 co.emo (%)	30.92	29.85	12.76

Table 3: Data Statistics. co.emo: refers to the percentage of co-existing emotions.

To pre-process the data, we utilised a tool designed for the specific characteristics of Twitter, i.e., misspellings and abbreviations (Baziotis et al., 2017). The tool offers different functionalities, such as tokenisation, normalisation, spelling correction, and segmentation. We used the tool to tokenise the text, convert words to lower case, normalise user mentions, urls and repeated-characters.

3.3 Multi-label Emotion Classification

We compared the performance of SpanEmo to some baseline as well as state-of-the-art models on all three languages. For experiments related to English, we selected seven models, while we chose three models for both Arabic and Spanish. We also include the results of BERT$_{\text{BASE}}$.

3.3.1 English

English models include JBNN (He and Xia, 2018), DATN (Yu et al., 2018), NTUA (Baziotis et al., 2018), RERc (Zhou et al., 2018), BERT$_{\text{BASE}}$+DK (Ying et al., 2019), BERT$_{\text{BASE}}$-GCN (Xu et al., 2020) and LEM (Fei et al., 2020). JBNN introduces a joint binary neural network, which focuses on learning the relations between emotions based on the theory of Plutchik's wheel of emotions (Plutchik, 1980), and then performing multi-label emotion classification via integrating

these label relations into the loss function. DATN proposes a dual attention transfer network to improve multi-label emotion classification with the help of sentiment classification, while NTUA is ranked the top-1 model of the SemEval2018 competition as it relies on different pre-training and fine-tuning strategies. RERc defines a ranking emotion relevant loss focused on incorporating emotion relations into the loss function to improve both emotion prediction and rankings of relevant emotions. Both BERT$_{\text{BASE}}$+DK and BERT$_{\text{BASE}}$-GCN utilise the same encoder as our own with the former considering additional domain knowledge (DK) and the latter capturing emotion relations through Graph Convolutional Network (GCN), respectively. LEM introduces a latent emotion memory network, in which the latent emotion module learns emotion distribution via a variational autoencoder, while the memory module captures features corresponding to each emotion.

3.3.2 Arabic

Arabic models consist of EMA (Badaro et al., 2018), TW-StAR (Mulki et al., 2018) and HEF (Alswaidan and Menai, 2020). EMA is the best performing model from the SemEval2018 competition on this set. It utilises various pre-processing steps (e.g. diacritics removal, normalisation, emojis transcription and stemming), as well as different classification algorithms. The TW-StAR model applies some pre-processing steps and then uses TF-IDF to learn features of a Support Vector Machine. HEF is based on a hybrid neural network, including different word embeddings (e.g. Word2Vec, Glove, FastText) plus variations of RNN neural networks, such as Long Short-Term Memory and Gated Recurrent Unit.

3.3.3 Spanish

Spanish models comprise TW-StAR (Mulki et al., 2018), ELiRF (González et al., 2018) and MI-LAB (Mohammad et al., 2018). The ELiRF model applies some pre-processing steps while also adapting the tweet tokeniser for Spanish tweets. MILAB is the best performing model from the SemEval2018 shared-task on this set.

4 Results

Table 4 presents the performance of our proposed approach (SpanEmo) on all three languages, in terms of micro F1-score (miF1), macro F1-score (maF1) and Jaccard index score (jacS), and compares it to the baseline and state-of-the-art models

---

[3] jacS is defined as the size of the intersection divided by the size of the union of the true label set and predicted label set.

discussed in section 3.3.

Language	English		
Model/Metric	miF1	maF1	jacS
JBNN	0.632	0.528	-
RERc	0.651	0.539	-
DATN	-	0.551	0.583
NTUA	0.701	0.528	0.588
$BERT_{base}$	0.695	0.520	0.570
$BERT_{base}$+DK	**0.713**	0.549	0.591
$BERT_{base}$-GCN	0.707	0.563	0.589
LEM	0.675	0.567	-
SpanEmo (ours)	**0.713**	**0.578**	**0.601**
	Arabic		
	miF1	maF1	jacS
Tw-StAR	0.597	0.446	0.465
EMA	0.618	0.461	0.489
$BERT_{base}$	0.650	0.477	0.523
HEF	0.631	0.502	0.512
SpanEmo (ours)	**0.666**	**0.521**	**0.548**
	Spanish		
	miF1	maF1	jacS
Tw-StAR	0.520	0.392	0.438
ELiRF	0.535	0.440	0.458
MILAB	0.558	0.407	0.469
$BERT_{base}$	0.596	0.474	0.487
SpanEmo (ours)	**0.641**	**0.532**	**0.532**

Table 4: The results of multi-label emotion classification on SemEval-2018 test set.

As shown in Table 4, our method outperformed all models on all languages, as well as on almost all metrics, with a marginal improvement of up to 1-1.3% for ENGLISH, 1.9-3.6% for ARABIC and 6.3-9.2% for SPANISH. This demonstrates the utility and advantages of SPANEMO, as well as the label-correlation aware loss for improving the performance of multi-label emotion classification in ENGLISH, ARABIC and SPANISH.

Based on the empirical results reported in Table 4, the following observations can be made. First, incorporating the relations between emotions into the models tends to lead to higher performance, especially for macro F1-score. For example, both DATN and LEM learn emotion-related features and achieve better performance than NTUA and BERT$_{BASE}$+DK. Additionally, ELiRF makes use of various sentiment/emotion features (i.e., learned from lexica) and it yielded the best performance

among the three compared models. This corroborates our earlier hypothesis that learning emotion-specific associations is crucial for improving the performance. Although BERT$_{BASE}$+DK adopts the same encoder as our own and adds domain knowledge, our method still performs strongly, especially for both macro F1- and jaccard score with a marginal improvement of up to 2.9% and 1%, respectively. In short, capturing emotion-specific associations as well as integrating the relations between emotions into the loss function, helped SPANEMO to achieve the best results compared with all models on almost all metrics.

4.1 Ablation Study

To understand the effect of our framework, we undertook an ablation study of the model performance under three settings: firstly, the model was trained only with BCE loss; secondly, it was trained only with LCA loss; and thirdly it was trained without the label segment. The third setting is equivalent to training the model as a simple multi-label classification task, by only considering the input sentence.

Language	English		
Model/Metric	miF1	maF1	jacS
SpanEmo (joint)	**0.713**	0.578	**0.601**
- $\mathcal{L}$ (LCA)	0.712	0.564	0.590
- $\mathcal{L}$ (BCE)	0.698	**0.583**	0.582
- Label Seg.	0.695	0.520	0.570
	Arabic		
	miF1	maF1	jacS
SpanEmo (joint)	**0.666**	0.521	**0.548**
- $\mathcal{L}$ (LCA)	0.654	0.481	0.534
- $\mathcal{L}$ (BCE)	0.660	**0.526**	0.532
- Label Seg.	0.650	0.477	0.523
	Spanish		
	miF1	maF1	jacS
SpanEmo (joint)	**0.641**	0.532	**0.532**
- $\mathcal{L}$ (LCA)	0.629	0.526	0.507
- $\mathcal{L}$ (BCE)	0.606	**0.544**	0.499
- Label Seg.	0.596	0.474	0.487

Table 5: Ablation experiment results. The second and third rows from each group correspond to the removal of the respective loss function, whereas the last row corresponding to the removal of the label segment.

Table 5 presents the results. When SPANEMO was trained without the LCA loss, the results dropped

Model/Metric	$\mathcal{L}$	miF1	maF1	jacS	miF1	maF1	jacS	miF1	maF1	jacS
English	$\mathcal{L}$		≥ 1 co.emo			≥ 2 co.emo			≥ 3 co.emo	
BERT$_{base}$	BCE	0.703	0.515	0.587	0.712	0.521	0.596	0.692	0.509	0.554
SpanEmo	BCE	0.716	0.563	0.599	0.737	0.578	0.629	0.748	0.597	0.639
SpanEmo	Joint	**0.724**	**0.590**	**0.613**	**0.746**	**0.606**	**0.648**	**0.753**	**0.624**	**0.643**
Arabic	$\mathcal{L}$		≥ 1 co.emo			≥ 2 co.emo			≥ 3 co.emo	
BERT$_{base}$	BCE	0.656	0.459	0.527	0.668	0.471	0.531	0.682	0.485	0.555
SpanEmo	BCE	**0.689**	0.518	**0.565**	0.709	0.536	0.586	0.745	0.567	**0.629**
SpanEmo	Joint	**0.689**	**0.534**	**0.565**	**0.710**	**0.551**	**0.587**	**0.746**	**0.584**	0.626
Spanish	$\mathcal{L}$		≥ 1 co.emo			≥ 2 co.emo			≥ 3 co.emo	
BERT$_{base}$	BCE	0.603	0.476	0.526	0.567	0.461	0.441	0.518	0.432	0.364
SpanEmo	BCE	0.653	0.528	0.561	0.646	0.528	0.519	**0.663**	0.566	**0.508**
SpanEmo	Joint	**0.662**	**0.565**	**0.581**	**0.655**	**0.568**	**0.530**	0.644	**0.570**	0.490

Table 6: Presenting the number of co-existing emotion classes. The best results in each language group are marked in bold.

by 1-2% for macro F1- and jaccard score. In addition, the results of SpanEmo dropped by 1-2% for two metrics apart from the macro F1-score when trained without the BCE loss. However, the removal of the label segment led to a much higher drop of 3-6%. The same patterns were also observed in the Arabic and Spanish experiments. This supports our earlier hypothesis that casting the task of multi-label emotion classification as span-prediction is beneficial for improving both the representation and performance of multi-label emotion classification.

5 Analysis

5.1 Prediction of Multiple Emotions

We additionally validated the effectiveness of our method for learning the multiple co-existing emotions on English, Arabic and Spanish sets. Table 6 presents the results, including BERT$_{BASE}$. SpanEmo demonstrated a strong ability to handle multi-label emotion classification much better than BERT$_{BASE}$. Since BERT$_{BASE}$ is trained only with binary cross-entropy (BCE) loss, here we include the results of our method trained only with this loss function. SpanEmo still achieved consistent improvement as the number of co-existing emotions increases, showing the usefulness of our method in learning multiple emotions. Improvement can clearly be observed for English and Arabic experiments, but not as much for Spanish. This may be

attributed to the high percentage of single-label data, which is around (40%) for Spanish, while it is lower than that for both English and Arabic. Obviously, SpanEmo can be used without LCA loss, and still obtain descent performance. Nevertheless, training our method jointly with the LCA loss leads to better results.

5.2 Learning Emotion-specific Associations

5.2.1 Word-Level

In this section, we present the top 10 words learned by SpanEmo for each emotion class by extracting the learned representations for each emotion class and all words in every input sentence, and then computing the similarity between them via cosine similarity. Finally, we performed this operation on all inputs in the SemEval2018 English validation set and then sorted all words for each emotion class in ascending order.

Table 7 presents the top-10 words per emotion class. As shown in Table 7, the words discovered by our framework are indicative of the corresponding emotion. This helps to show that SpanEmo learns meaningful associations between emotion classes and words automatically, which can be beneficial for feature extraction and learning. Additionally, SpanEmo demonstrated that it can learn diverse words as well as shareable words across some emotions. For example, the words $\{pissed, wrath, smashed\}$ are associated with both anger and disgust, demonstrating the ability of SpanEmo to learn the relations between emotions.

Emotion	Top 10 Words
anger	anger pissed wrath idiots dammit kicking irritated thrown smashed complain
anticipation	prediction planning mailsport assumptions upcoming waiting route waited frown ideas
disgust	disgusting smashed gross hate pissed wrath dirty awful vile dumb
fear	nervous fear terror frightening afraid frown panic terrifying scary dreading
joy	happy excitement joyful congratulations glad delightful excited adorable amusing smiling
love	love sweetness loved hug mate lucky carefree shine care gracious
optimism	optimism integrity salvation persevere perspective bright effort faith glad lord
pessimism	hopeless frown disappointed weary dread despair depressing chronic suicide pain
sadness	sadness frown depressing saddened hurt disappointed weary upset sorrow hate
surprise	stunned awestruck shocking awe mailsport buster genuinely curious hardly believing
trust	integrity shine respect courage sign effort confident faith easy kindness

Table 7: Top 10 words associated with each corresponding emotion.
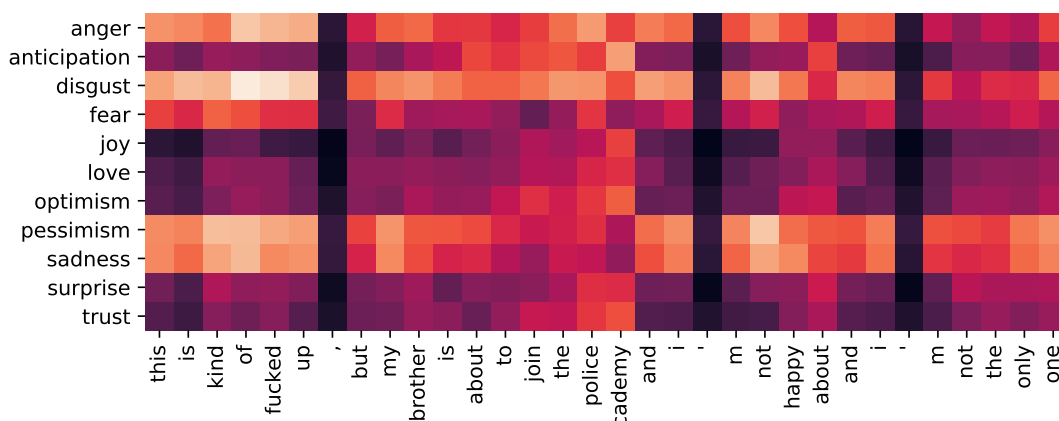
Figure 2: Visualisation on an example. The left presents the emotion labels, and the bottom presents the example. Each cell shows the cosine similarity value computed via using the hidden representation of each word and label. Lighter colour indicates higher similarity.

5.2.2 Sentence-Level

We visualised an example from the ENGLISH validation set annotated with four emotions, which were anger, disgust, pessimism and sadness. Our goal was to determine whether by adding emotion classes to the example, SpanEmo could learn their associations to each other (i.e., associations between emotion classes and words in the example). To compute the similarity between emotion classes and words in the example, we also followed the same process discussed in section 5.2.1.

Figure 2 presents the results, where lighter cells indicate higher similarity, while darker cells indicate lower similarity. As shown in Figure 2, the learned representations capture the association between the correct emotion label set and every token in the example. Interestingly, we can also observe that the word "happy" is usually expressed as a positive emotion, but, in this context, this word becomes negative and the model learns this contextual information. Moreover, the phrase "about to join the police academy" is associated with "anticipation", which makes sense although this class is not part of the correct label set. This demonstrates the utility and advantages of our approach not only in deriving associations reported in the annotations, but also providing us with a mechanism to explore additional information beyond them.
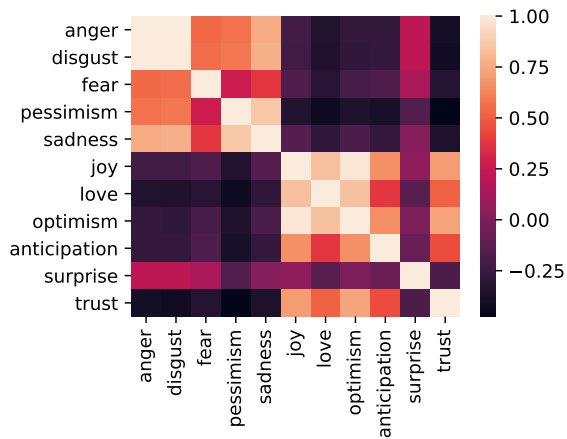
5.3 Label Correlations

Since one of the research questions in this paper was to learn the multiple co-existing emotions from a multi-label emotion data set, we analysed the learned emotion correlations from SpanEmo and compared them to those adopted from the ground truth labels in the SemEval2018 validation set. Figure 3 presents the two emotion correlations as obtained from the ground truth labels and from the predicted labels, respectively.

It can be observed that Figure 3(b) is almost identical to 3(a), demonstrating that our method in capturing the emotion correlations is in line with
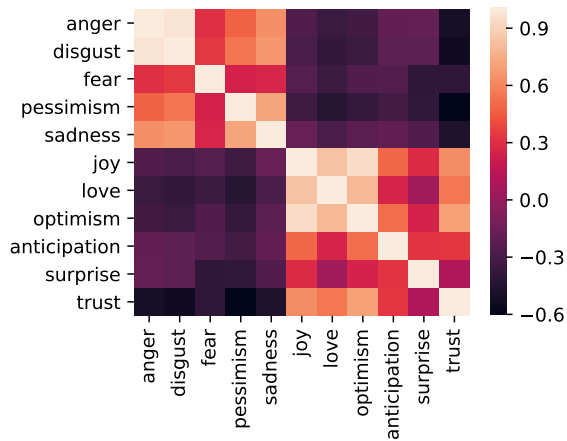
(a) Emotion correlations: GT

(b) Emotion correlations: Prediction

Figure 3: The top plot presents emotion correlations obtained from the ground truth (GT) labels, whereas the bottom plot presents emotion correlations obtained from the predicted labels.

what the emotion annotations have revealed. 3(b), which was learned by SpanEmo, also highlights that negative emotions are positively correlated with each other, and negatively correlated with positive emotions. For example, "anger and disgust" share almost the same patterns, which is consistent with the studies of Mohammad and Bravo-Marquez (2017) and Agrawal et al. (2018), both of which report the same issue with negative emotions of "anger" and "disgust", as they are easily confused with each other. This is not surprising as their manifestation in language is quite similar in terms of the use of similar words/expression. We also noted this finding when analysing the top-10 key words learned by SpanEmo in section 5.2.1. In short, taking into account emotion correlations is crucial for multi-label emotion classification in addressing the ambiguity characteristic of the task, especially for

emotions that are highly correlated.

5.4 Influence of Parameter (α)

SpanEmo was trained with BCE loss and with LCA loss via a weight (α), whose impact on the results is presented in Figure 4. It should be mentioned that this analysis was performed on the validation set of SemEval2018 data set. The lower bound (i.e., 0.0) indicates that the model was trained only with the BCE loss, whereas the upper bound (i.e., 1.0) indicates that it was trained only with the LCA loss. When the value of α increased from 0.0 to 0.5, the results first improved considerably and then gradually deteriorated apart from the results of the macro F1-score. The results of BCE loss favoured the micro F1- and jaccard score, whereas the results of LCA loss favoured the macro F1-score. However, integrating LCA with BCE can balance the results across all three metrics, resulting in strong performance. The best results were achieved on almost all metrics when the value of α was set to 0.2. Thus, we set the value of parameter α to 0.2 for all experiments reported in this paper.
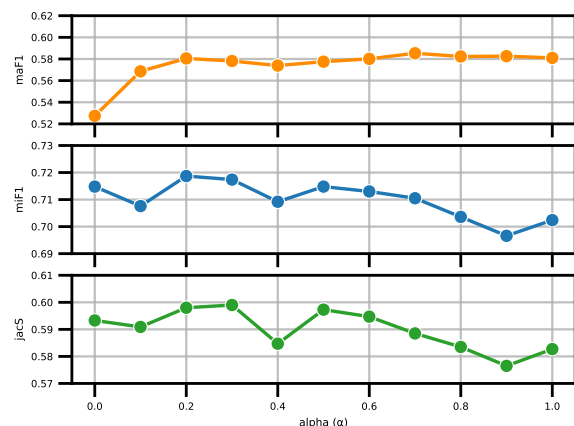

Figure 4: Sensitivity analysis of parameter (α). Note that $\alpha = 0.0$ means that only BCE loss is used in training SpanEmo, whereas $\alpha = 1.0$ means that only LCA loss is utilised in training it.

6 Related Work

There is a large body of NLP literature on emotion recognition (Mohammad and Alm, 2015). Earlier studies focused on lexicon-based approaches, which make use of some words and their corresponding labels to identify emotions in text, e.g. NRC[4] (Mohammad and Turney, 2013) and

---

[4]Bravo-Marquez et al. (2016) proposes an approach for expanding it for the language used in Twitter.

EMOSENTICNET (Poria et al., 2014). Other methods treat the emotion recognition task as a supervised learning task, in which a learner (e.g. linear classifier based methods) is trained on the features of labelled data to classify inputs into one label (Bostan and Klinger, 2018; Liew et al., 2016; Mohammad et al., 2015; Tang et al., 2013; Wang et al., 2012; Aman and Szpakowicz, 2007).

More recently, several neural network models have been developed for this task, obtaining competitive results on different emotion data sets. Some of these models generally focus on a single-label emotion classification, in which only a single label is assigned to each input (Islam et al., 2019; Xia and Ding, 2019; Alhuzali et al., 2018b,a; Agrawal et al., 2018; Saravia et al., 2018; Felbo et al., 2017; Abdul-Mageed and Ungar, 2017). Other models have also been proposed for multi-label emotion classification, in which one or more labels are assigned to each input (see detailed description in section 3.3).

Our work is motivated by research focused on learning features corresponding to each emotion as well as incorporating the relations between emotions into a loss function (Fei et al., 2020; He and Xia, 2018). Our work differs from these two works in the following ways: i) our method learns features related to each corresponding emotion without relying on any external resources (e.g. lexicons). ii) We further integrated the relations between emotions into the loss function by taking advantage of the label co-occurrences in a multi-label emotion data set. In this respect, our approach does not rely on any theory of emotion. iii) We empirically evaluated our method for three languages, demonstrating its effectiveness as being language agnostic. In contrast to previous research, we focus on both learning emotion-specific associations and integrating the relations between emotions into the loss function.

7 Conclusion

We have proposed a novel framework "SpanEmo" aimed at casting multi-label emotion classification as a span-prediction problem. We demonstrated that our proposed method outperforms prior approaches reported in the literature on three languages (i.e., English, Arabic and Spanish). Our empirical evaluation and analyses also demonstrated the utility and advantages of our method for multi-label emotion classification, specifically the addi-

tion of emotion classes to the input sentence, which helped the model learn emotion-specific associations and increase its performance. Finally, training our method with LCA loss jointly led to better results, showing the benefits of integrating the relations between emotions into the loss function.

The standard approach in a multi-label emotion classification problem often focuses on modelling individual emotion independently. In this respect, most existing methods do not take into account label dependencies while learning emotion-specific associations. However, we demonstrated the effectiveness of including label information to the input sentence when training SpanEmo, helping it achieve better performance and capture emotion correlations. We hope that this study will inspire the community to investigate further the vital role of learning label dependencies and associations corresponding to each emotion.

References

Muhammad Abdul-Mageed and Lyle Ungar. 2017. EmoNet: Fine-Grained Emotion Detection with Gated Recurrent Neural Networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 718–728.

Ameeta Agrawal, Aijun An, and Manos Papagelis. 2018. Learning emotion-enriched word representations. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 950–961.

Yaara Benger Alaluf and Eva Illouz. 2019. Emotions in consumer studies. *The Oxford Handbook of Consumption*, page 239.

Hassan Alhuzali, Muhammad Abdul-Mageed, and Lyle Ungar. 2018a. Enabling deep learning of emotion with first-person seed expressions. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 25–35.

Hassan Alhuzali and Sophia Ananiadou. 2019. Improving classification of adverse drug reactions through using sentiment analysis and transfer learning. In *Proceedings of the 18th BioNLP Workshop and*

Shared Task, pages 339–347, Florence, Italy. Association for Computational Linguistics.

Hassan Alhuzali, Mohamed Elaraby, and Muhammad Abdul-Mageed. 2018b. UBC-NLP at IEST 2018: Learning implicit emotion with an ensemble of language models. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 342–347, Brussels, Belgium. Association for Computational Linguistics.

Nourah Alswaidan and Mohamed El Bachir Menai. 2020. Hybrid feature model for emotion recognition in arabic text. *IEEE Access*, 8:37843–37854.

Saima Aman and Stan Szpakowicz. 2007. Identifying expressions of emotion in text. In *International Conference on Text, Speech and Dialogue*, pages 196–205. Springer.

Mario Ezra Aragón, Adrian Pastor López-Monroy, Luis Carlos González-Gurrola, and Manuel Montes. 2019. Detecting depression in social media using fine-grained emotions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1481–1486.

Gilbert Badaro, Obeida El Jundi, Alaa Khaddaj, Alaa Maarouf, Raslan Kain, Hazem Hajj, and Wassim El-Hajj. 2018. EMA at SemEval-2018 task 1: Emotion mining for Arabic. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 236–244, New Orleans, Louisiana. Association for Computational Linguistics.

Christos Baziotis, Athanasiou Nikolaos, Alexandra Chronopoulou, Athanasia Kolovou, Georgios Paraskevopoulos, Nikolaos Ellinas, Shrikanth Narayanan, and Alexandros Potamianos. 2018. Ntua-slp at semeval-2018 task 1: Predicting affective content in tweets with deep attentive rnns and transfer learning. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 245–255.

Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. 2017. Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 747–754.

Laura-Ana-Maria Bostan and Roman Klinger. 2018. An Analysis of Annotated Corpora for Emotion Classification in Text. In *Proceedings ofthe 27th International Conference on Computational Linguistics*, pages 2104–2119.

Felipe Bravo-Marquez, Eibe Frank, Saif M Mohammad, and Bernhard Pfahringer. 2016. Determining word-emotion associations from tweets by multilabel classification. In *2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 536–539. IEEE.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *to appear in PML4DC at ICLR 2020*.

Xuetong Chen, Martin D Sykora, Thomas W Jackson, and Suzanne Elayan. 2018. What about mood swings: Identifying depression on twitter with temporal measures of emotions. In *Companion Proceedings of the The Web Conference 2018*, pages 1653–1660. International World Wide Web Conferences Steering Committee.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Hao Fei, Yue Zhang, Yafeng Ren, and Donghong Ji. 2020. Latent emotion memory for multi-label emotion classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7692–7699.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1615–1625.

José-Ángel González, Lluís-F. Hurtado, and Ferran Pla. 2018. ELiRF-UPV at SemEval-2018 tasks 1 and 3: Affect and irony detection in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 565–569, New Orleans, Louisiana. Association for Computational Linguistics.

Huihui He and Rui Xia. 2018. Joint binary neural network for multi-label learning with applications to emotion classification. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 250–259. Springer.

Jonathan Herzig, Guy Feigenblat, Michal Shmueli-Scheuer, David Konopnicki, and Anat Rafaeli. 2016. Predicting customer satisfaction in customer support conversations in social media using affective features. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, pages 115–119.

Jumayel Islam, Robert E. Mercer, and Lu Xiao. 2019. Multi-channel convolutional neural network for twitter emotion and sentiment recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*

1582

(Long and Short Papers), pages 1355–1365, Minneapolis, Minnesota. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Jasy Liew, Suet Yan, and Howard R Turtle. 2016. Exploring Fine-Grained Emotion Detection in Tweets. In *The 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 73–80.

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 and Cecilia Ovesdotter Alm. 2015. Computational analysis of affect and emotion in language. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*.

Saif M. Mohammad and Felipe Bravo-Marquez. 2017. Emotion intensities in tweets. In *Proceedings of the sixth joint conference on lexical and computational semantics (*Sem)*, Vancouver, Canada.

Saif M Mohammad and Svetlana Kiritchenko. 2013. Using nuances of emotion to identify personality. *arXiv preprint arXiv:1309.6352*.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. 29(3):436–465.

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.

Hala Mulki, Chedi Bechikh Ali, Hatem Haddad, and Ismail Babaoğlu. 2018. Tw-StAR at SemEval-2018 task 1: Preprocessing impact on multi-label emotion classification. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 167–171, New Orleans, Louisiana. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch.

Robert Plutchik. 1980. *Emotion: A psychoevolutionary synthesis*. Harpercollins College Division.

Robert Plutchik. 1984. Emotions: A general psychoevolutionary theory. *Approaches to emotion*, 1984:197–219.

Soujanya Poria, Alexander Gelbukh, Erik Cambria, Amir Hussain, and Guang-Bin Huang. 2014. Emosenticspace: A novel framework for affective common-sense reasoning. *Knowledge-Based Systems*, 69:108–123.

Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. 2020. Kuisail at semeval-2020 task 12: Bert-cnn for offensive speech identification in social media.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized Affect Representations for Emotion Recognition. In *Empirical Methods in Natural Language Processing*, pages 3687–3697.

Duyu Tang, Bing Qin, Ting Liu, and Zhenghua Li. 2013. Learning sentence representation for emotion classification on microblogs. In *Natural Language Processing and Chinese Computing*, pages 212–223. Springer.

Svitlana Volkova and Yoram Bachrach. 2016. Inferring perceived demographics from user emotional tone and user-environment emotional contrast. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1567–1578.

Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P Sheth. 2012. Harnessing twitter" big data" for automatic emotion identification. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pages 587–592. IEEE.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Rui Xia and Zixiang Ding. 2019. Emotion-cause pair extraction: A new task to emotion analysis in texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1003–1012, Florence, Italy. Association for Computational Linguistics.

Peng Xu, Zihan Liu, Genta Indra Winata, Zhaojiang Lin, and Pascale Fung. 2020. Emograph: Capturing emotion correlations using graph networks. *arXiv preprint arXiv:2008.09378*.

Pengcheng Yang, Xu Sun, Wei Li, Shuming Ma, Wei Wu, and Houfeng Wang. 2018. Sgm: Sequence generation model for multi-label classification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3915–3926.

Chih-Kuan Yeh, Wei-Chieh Wu, Wei-Jen Ko, and Yu-Chiang Frank Wang. 2017. Learning deep latent spaces for multi-label classification. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 2838–2844.

Wenhao Ying, Rong Xiang, and Qin Lu. 2019. Improving multi-label emotion classification by integrating both general and domain-specific knowledge. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 316–321, Hong Kong, China. Association for Computational Linguistics.

Jianfei Yu, Luis Marujo, Jing Jiang, Pradeep Karuturi, and William Brendel. 2018. Improving multi-label emotion classification via sentiment classification with dual attention transfer network. ACL.

Min-Ling Zhang and Zhi-Hua Zhou. 2006. Multilabel neural networks with applications to functional genomics and text categorization. *IEEE transactions on Knowledge and Data Engineering*, 18(10):1338–1351.

Deyu Zhou, Yang Yang, and Yulan He. 2018. Relevant emotion ranking from text constrained with emotion relationships. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 561–571.