

Article

Mixing and Matching Emotion Frameworks: Investigating Cross-Framework Transfer Learning for Dutch Emotion Detection

Luna De Bruyne , Orphée De Clercq  and Véronique Hoste 

Language and Translation Technology Team, Ghent University, 9000 Ghent, Belgium; orphee.declercq@ugent.be (O.D.C.); veronique.hoste@ugent.be (V.H.)

* Correspondence: luna.debruyne@ugent.be; Tel.: +32-9-331-1938

Abstract: Emotion detection has become a growing field of study, especially seeing its broad application potential. Research usually focuses on emotion classification, but performance tends to be rather low, especially when dealing with more advanced emotion categories that are tailored to specific tasks and domains. Therefore, we propose the use of the dimensional emotion representations valence, arousal and dominance (VAD), in an emotion regression task. Firstly, we hypothesize that they can improve performance of the classification task, and secondly, they might be used as a pivot mechanism to map towards any given emotion framework, which allows tailoring emotion frameworks to specific applications. In this paper, we examine three cross-framework transfer methodologies: multi-task learning, in which VAD regression and classification are learned simultaneously; meta-learning, where VAD regression and emotion classification are learned separately and predictions are jointly used as input for a meta-learner; and a pivot mechanism, which converts the predictions of the VAD model to emotion classes. We show that dimensional representations can indeed boost performance for emotion classification, especially in the meta-learning setting (up to 7% macro F1-score compared to regular emotion classification). The pivot method was not able to compete with the base model, but further inspection suggests that it could be efficient, provided that the VAD regression model is further improved.

Keywords: emotion detection; multi-task learning; transfer learning; emotion frameworks



Citation: De Bruyne, L.; De Clercq, O.; Hoste, V. Mixing and Matching Emotion Frameworks: Investigating Cross-Framework Transfer Learning for Dutch Emotion Detection. *Electronics* **2021**, *10*, 2643. <https://doi.org/10.3390/electronics10212643>

Academic Editor: Miguel A. Alonso

Received: 30 September 2021

Accepted: 27 October 2021

Published: 29 October 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Since the year 2000, sentiment analysis is one of the most well-studied research domains in natural language processing (NLP), not in the least because of its broad application potential. Companies and organisations use it to learn more about (potential) customers or clients [1] or to gauge their online reputation [2]. Traditionally, sentiment analysis focused on the study of polarity with the goal of classifying textual instances as either positive or negative [1]. However, from a company perspective, it can be more interesting to pinpoint exactly what customers are talking about: for example, if they mention a product in an online review, they can be very specific as to which aspects they liked (e.g., quality and design) and/or disliked (e.g., user-friendliness). This had led to the emergence of aspect-based sentiment analysis, which focuses on the detection of sentiment expressions on the aspect or feature level [3].

In more recent years, the field advanced from analysing polarity to detecting more fine-grained emotions [4]. The goal in emotion analysis is to analyse specific emotional states such as *anger*, *joy* and *sadness* or emotional dimensions like *valence* and *arousal*. By studying emotions, companies get more hands-on insights into which customer responses require an immediate action. For example, understanding when a customer is clearly angry or sad is more insightful than the label *negative* in the framework of customer response management

or when thinking about employing chatbots for customer support. Moreover, the emotions of interest might vary regarding the application or domain.

When dealing with emotion analysis for specific applications, this leads to an important methodological issue, namely the choice of emotion framework. Many studies focus on the classification of textual utterances into basic emotion categories following the frameworks of Ekman (*anger, disgust, fear, joy, sadness and surprise*) [5] and Plutchik (*anger, anticipation, disgust, fear, joy, sadness, surprise and trust*) [6]. However, multiple researchers have emphasized the need of studying emotions not only in terms of basic emotion categories, but based on emotional dimensions like *valence, arousal and dominance* (VAD) as well [7,8].

In earlier work, we have already criticized the apparent arbitrariness with which an emotion framework is chosen for studies in NLP [9]. Mostly, a data-driven motivation or experimentally grounded choice is lacking. However, some researchers see benefits in tailoring the emotion label set to the task at hand. In the case of crisis communication, for example, it would be appropriate to employ the crisis-related emotion framework of Jin et al. [10], as proposed by Hoste et al. [11].

Although the emotional nuances in different label sets could be useful, tailoring these sets to specific applications or domains might introduce different challenges: (a) resources will need to be created for every specific application and domain, (b) emotion detection resources will be scattered over different frameworks, and (c) emotion detection systems will not be generalizable.

Cross-framework transfer learning methods could mitigate these challenges. Fine-tuning pre-trained models, multi-task learning or label space mapping methods can considerably decrease the amount of required training data, as this allows for the transfer of knowledge across divergent emotion frameworks.

A straight-forward approach to shift between frameworks is to map discrete categories into a three-dimensional space, which corresponds to Mehrabian and Russell's claim that all affective states can be represented by the dimensions *valence, arousal and dominance* [12]. This mapping to and from the VAD space can be regarded as a pivot mechanism. Regardless of the specific mapping technique (e.g., linear regression, kNN or lexicon-based mappings), this idea opens possibilities. Given an accurate mapping technique and a well-performing emotion analysis system that predicts values for *valence, arousal and dominance*, the predicted VAD values can be converted to any categorical emotion label set. Emotion frameworks can then easily be tailored to specific tasks and domains, broadening their scope of application in e.g., customer service management or conversational agents.

Moreover, previous experiments for Dutch emotion detection revealed that the classification of emotional categories (*anger, fear, joy, love and sadness*) is very challenging, while more promising results were found for VAD regression [13]. Transferring information from the regression task to improve performance on the classification task would therefore be an interesting line of research.

This study investigates the potential of dimensional representations and revolves around two research questions: (a) can dimensional representations serve as an aid in the prediction of emotion categories and (b) can dimensional representations contribute in tailoring label sets to specific tasks and domains?

Our research focuses on Dutch emotion detection and will make use of the EmotioNL dataset [13]. We examine three cross-framework transfer methodologies, namely multi-task learning, meta-learning and the aforementioned pivot mechanism. In the multi-task setting, the VAD regression task and classification task are learned simultaneously. In the meta-learner approach, two systems are trained separately, one for VAD regression and one for emotion classification. We will investigate whether a multi-task learner or a meta-learner that exploits both sources of information is favorable compared to a system that only uses one source. These models will be compared to a system relying on a pivot method, using solely dimensional representations. The code is publicly available at

<https://github.com/LunaDeBruyne/Mixing-Matching-Emotion-Frameworks> (accessed on 30 September 2021).

We thus contribute to the field of emotion analysis in NLP by leveraging dimensional representations to increase the performance of emotion classification and by proposing a method to tailor label sets to specific applications.

The remainder of this paper is organised as follows: in Section 2, related work on the combination of categorical and dimensional frameworks in emotion detection is discussed. Section 3 describes the materials and methods of our study and gives an overview of the used data (Section 3.1) and a description of the experimental setup (Section 3.2). Results are reported in Section 4 and further discussed in Section 5. This paper ends with a conclusion in Section 6.

2. Related Work

Our previous work on Dutch emotion detection focused on the prediction of the classes *joy, love, anger, fear, sadness* or *neutral* and the emotional dimensions *valence, arousal* and *dominance* in Dutch Twitter messages and captions from reality TV-shows [13]. We found that the classification results were low (54% accuracy for tweets and 48% for captions). However, the results for emotional dimensions were more promising (0.64 Pearson's r for both domains). This observation, together with the issue of having specialised categorical labels for specific tasks/domains, reinforces the urgency to focus more on dimensional models and investigate their potential of aiding emotion classification by means of transfer learning.

Multi-task learning settings have proven successful in many tasks related to emotion and sentiment analysis [14,15]. Although there are not many studies that perform transfer learning with multiple emotion frameworks, there are various studies that employ multi-task learning by jointly training emotion detection with sentiment analysis [16,17] or other related tasks [18]. All of these studies suggest that multi-task frameworks outperform single-task experiments and thus motivate the idea to train emotion classification and VAD regression jointly, especially as VAD probably contains more valuable emotional information than sentiment (which only contains the first dimension: *valence*).

Various studies have also investigated how to deal with disparate label spaces. Mostly, this involves a mapping between categorical and dimensional frameworks, e.g., in the work of Stevenson et al. [19] and Buechel and Hahn [20,21]. In these studies, scores for *valence, arousal* and *dominance* were used to predict intensity values for the basic emotion categories *happiness, anger, sadness, fear* and *disgust*, and vice versa. To this end, linear regression [19], a kNN model [20] and a multi-task feed-forward network [21] were used. Especially this last method provided promising results, where a Pearson correlation of 0.877 was obtained for mapping dimensions to categories and 0.853 for the other direction.

A straightforward approach is to map discrete categories directly into the VAD space, which corresponds to Mehrabian and Russell's claim that all affective states can be represented by the dimensions *valence, arousal* and *dominance* [12]. Figure 1 shows the positions of Ekman's basic emotions in the VAD space, based on the scores of these terms in Mehrabian and Russell [12]. Calvo and Mac Kim employ this idea and apply it directly to the task of emotion detection [22]. They obtain lexicon scores for emotion words related to the categories *anger/disgust, fear, joy* and *sadness* by looking them up in the Affective Norms for English Words (ANEW) [23], and map the center of each of these categories in the VAD space. Then, they calculate VAD scores for sentences (again using the ANEW lexicon), which are placed in the emotional space as well. By computing cosine similarity between the sentence and the previously mapped emotion categories, the emotional category of the sentence can be determined. This lexicon-based mapping approach has as an advantage that no annotated categories are needed, in contrast to the previously discussed approaches which do require annotated categories to learn a mapping.

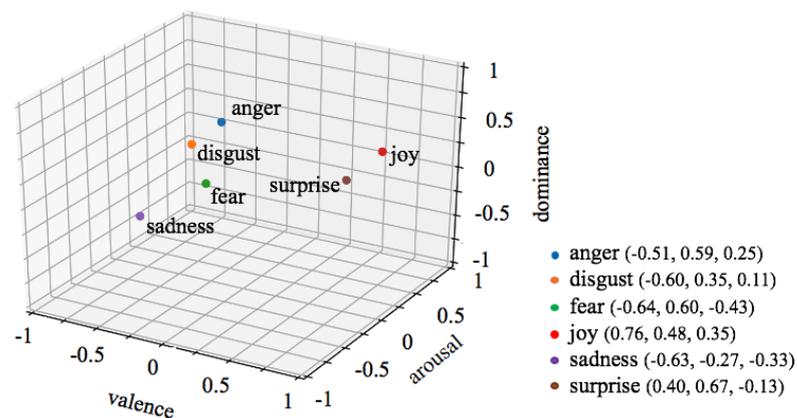


Figure 1. Mapping of Ekman's six into the VAD-space, figure based on the scores for the English Ekman terms of Mehrabian and Russell [12].

Besides mapping between emotion frameworks, a similar line of research deals with the unification of disparate label spaces in emotion and sentiment resources. Examples of merging sentiment lexica are [24–26] for emotion lexica and [27] for emotion datasets. Techniques exist out of Bayesian models [24], variational autoencoders [25,26] and rule-based combination techniques [27] to map lexica or datasets with different labels into the same space.

3. Materials and Methods

In this section, we describe the data and experimental setup to thoroughly investigate the potential of dimensional representations in (a) improving emotion classification, and (b) tailoring the label set to specific tasks and domains by mapping emotional dimensions to categories.

3.1. Data

For this study, the EmotioNL dataset is used [13]. This dataset consists of Dutch data in two domains: Twitter posts (Tweets subcorpus) and utterances from reality TV-shows (Captions subcorpus). The Tweets subcorpus consists of 1000 tweets that all contain at least one out of a list of 72 emojis. The Captions subcorpus consists of 1000 utterances from transcriptions of three emotionally loaded Flemish reality TV-shows (*Blind getrouwd*; *Bloed, zweet en luxeproblemen*; and *Ooit vrij*), more or less equally distributed over the shows (335 instances from *Blind getrouwd*, 331 from *Bloed, zweet en luxeproblemen* and 334 from *Ooit vrij*).

All data were annotated with both categorical labels and dimensions. For the categorical annotation, the instances were labeled with one out of six labels: *joy*, *love*, *anger*, *fear*, *sadness*, or *neutral*. The dimensional annotations are real-valued scores from 0 to 1 for the dimensions *valence*, *arousal* and *dominance*. An annotated example of one instance per domain is shown in Table 1.

Table 1. Text examples from the Tweets and Captions subcorpora with their assigned categorical and dimensional label (V = valence, A = arousal, D = dominance).

Corpus	Text Example	Categorical	Dimensional		
			V	A	D
Tweets	Vanmorgen vroeg opgestaan en de zon schijnt al lekker volop ☺ Vandaag er even op uit en genieten van de zon. Fijne dag allemaal <i>EN: Woke up early this morning and the sun is already shining brightly ☺ Going out today to enjoy the sun. Have a nice day everyone</i>	joy	0.819	0.578	0.523
Captions	Gij komt hier altijd met van die stomme flauwekul, gj. Kheb da nie nodig. <i>EN: You always come here with that stupid bullshit. I don't need that.</i>	anger	0.071	0.713	0.649

3.2. Experimental Setup

To answer the first research question, we will look at two ways to leverage dimensional representations in aiding emotion classification, namely multi-task learning and a stacking-based meta-learning method. The second question, namely whether dimensions can be mapped to categories to tailor label sets to specific applications, will be investigated by means of a pivot method, where we employ predictions from a dimensional model to predict emotion classes. Each of these models together with a baseline model are described in closer detail in the following sections.

3.2.1. Base Model: RobBERT

We will employ the Dutch transformer model RobBERT [28], the Dutch version of the robustly optimized RoBERTa [29], which is trained on 39 GB of common crawl data [30]. It consists of 12 self-attention layers with 12 heads, and has 117 M trainable parameters. Previous experiments showed that this model achieves the best performance for emotion detection [13] in comparison to the BERT-based BERTje model [31].

We implement the model using HuggingFace's Transformers library [32]. The fine-tuning process uses AdamW optimizer [33] and the ReduceLROnPlateau learning rate scheduler with learning rate $5e - 5$. The loss function is Binary Cross Entropy for the classification task and Mean Squared Error loss for regression. We set dropout to 0.2 and use GELU as activation function in the implementation of [34]. The maximum sequence length is 64 tokens. The model is trained for 5 epochs for classification and 10 for regression with a batch size of 64 on an Nvidia Tesla V100 GPU. As we are dealing with small datasets (1000 instances per domain and task), the model is evaluated using 10-fold cross-validation.

3.2.2. Multi-Task Learning

In this setting, the classification (categories) and regression (dimensions) models are trained simultaneously (see Figure 2). We use the same architecture and hyperparameters as in the base model. The RobBERT feature encoder allows for hard parameter sharing where the learning of features for the emotion classes and VAD prediction happens simultaneously, but has separate task-specific output layers. The losses (Binary Cross Entropy for emotion classification and Mean Squared Error loss for VAD) are averaged according to pre-defined weights. We test three different ratios: one where VAD and classification are weighed equally (both 0.5), one where classification outweighs VAD (0.75 for classification and 0.25 for VAD) and one where VAD has the largest weight (0.75 for VAD and 0.25 for classification).

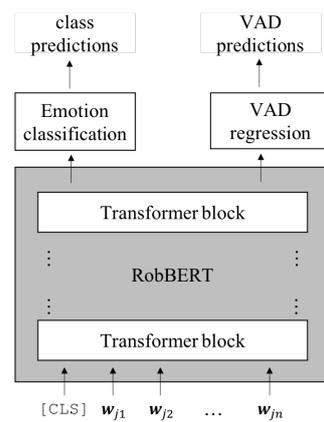


Figure 2. Schematic representation of the multi-task learning architecture.

3.2.3. Meta-Learner

The meta-learner approach is another way of leveraging the information in dimensional representations as these are combined with categorical inputs. However, in this setting, no parameters are shared between the tasks. Instead, a stacking ensemble is used in which two base models are trained, one for VAD regression and one for emotion classification. The predictions (or probabilities in the case of classification) are concatenated (six values for classification and three values for VAD) and used as input for a meta-learner algorithm, in this case a support vector machine for classification and a linear regression model for VAD. A diagram of the proposed architecture is depicted in Figure 3.

Nested cross-validation is used for this approach: when training the base models, for every test fold, a model is trained on eight folds and predictions are saved on the remaining fold. This is repeated, so that, for every test fold, predictions for the other nine folds have been made. These predictions will be used in the training phase of the meta-learner. Afterwards, the model is trained again using regular cross-validation, in order to save predictions for the test fold, based on training on nine folds. These predictions are then used during the test phase of the meta-learner.

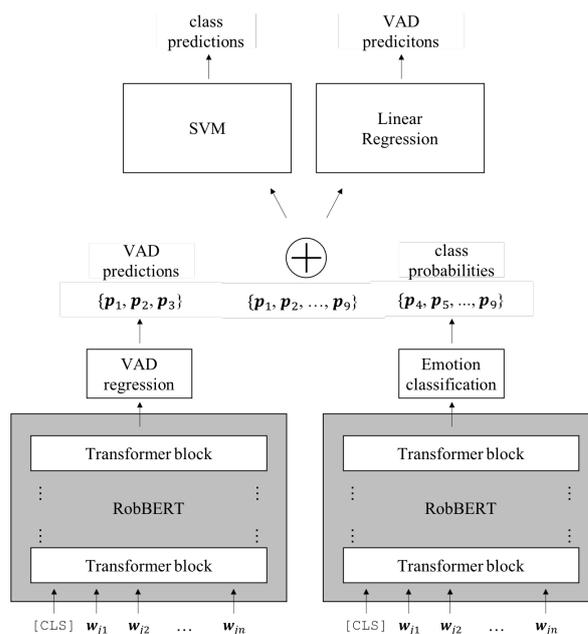


Figure 3. Schematic representation of the meta-learning architecture.

The support vector machine is trained with default settings in the Scikit-learn library [35]: we use a linear kernel and 1.0 as regularization parameter C. Hinge loss and L2 penalty is used for classification. The balanced mode is used, meaning that class weights are taken into account, inversely proportional to class frequencies.

3.2.4. Pivot Method

Contrary to the two previous approaches, the pivot method makes only use of the VAD annotations instead of both the dimensional and categorical data. It starts from predicting VAD scores through a transformer model, and these predictions are then transformed to classes by means of a rule-based mapping (see Figure 4). Although several mapping techniques have been investigated in related work (see Section 2), these approaches are not eligible for a pivot method, as they all rely on data in a bi-representational format and thus also require categorical data for a mapping to be learned. However, the idea of a pivot is to be able to map to any possible label set, without having to rely on any annotations for those labels.

The rule-based mapping works as follows: we look up the emotion terms from our label set (*anger*, *fear*, *joy*, *love* and *sadness*) in the definition list with VAD scores of Mehrabian and Russell [12] and scale them to a range from 0 to 1 to match the VAD annotation framework of the dataset. The scores can be found in Table 2. Following [22], we place both the textual instances to be classified and the vectors for the categorical emotion terms in the three-dimensional space.

We start by drawing some general rules for *anger*, *fear*, *joy* and *sadness*, as shown in Table 3 (at this point, *love* and *neutral* are not taken into consideration). If a class cannot be matched based on these rules, then we calculate cosine distance between the instance that needs to be classified and each emotion class vector (here *love* and *neutral* are included, the last one being defined as {0.5, 0.5, 0.5}). The class which has the smallest cosine distance to the instance is then assigned.

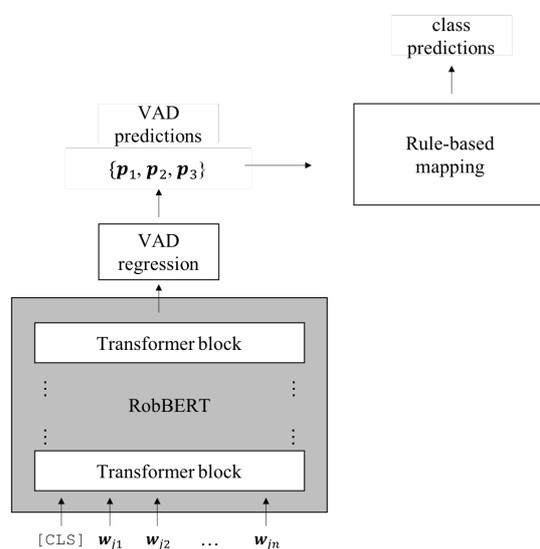


Figure 4. Schematic representation of the pivot method.

Table 2. Scores for valence, arousal and dominance according to the definitions of [12], scaled to a range from 0 to 1.

	V	A	D
Anger	0.245	0.795	0.625
Fear	0.180	0.800	0.285
Joy	0.905	0.755	0.730
Love	0.910	0.825	0.475
Sadness	0.185	0.365	0.335

Table 3. Mapping rule used in the pivot method.

if $V < 0.5$ and $A > 0.5$ and $D > 0.5$:
$class \leftarrow anger$
elif $V < 0.5$ and $A > 0.5$ and $D < 0.5$:
$class \leftarrow fear$
elif $V > 0.5$ and $A > 0.5$ and $D < 0.5$:
$class \leftarrow joy$
elif $V < 0.5$ and $A < 0.5$ and $D < 0.5$:
$class \leftarrow sadness$
else:
<i>Find class with smallest cosine distance</i>

3.2.5. Evaluation

Our experiments will be evaluated using three metrics: macro-averaged F1 (F1), micro-averaged F1 (Acc.) and cost-corrected accuracy (CC-Acc.). Cost-corrected accuracy is similar to normal accuracy, but, here, a cost matrix with specific weights is taken into account [13]. This way, misclassifications within the correct polarity are punished less than misclassifications in the opposite polarity (e.g., misclassifying an instance of *fear* as *sadness* has a lower weight than misclassifying *love* as *anger*).

4. Results

We report results for the three metrics (macro F1, accuracy and cost-corrected accuracy) for the base transformer model, the multi-task model in its three settings (equal weights, higher weight for classification and higher weight for regression), the meta-learner and the pivot model. The results for Tweets are shown in Table 4 for categories and Table 5 for VAD, while results for Captions are shown in Tables 6 and 7.

Table 4. Macro F1, accuracy and cost-corrected accuracy for the different models on the classification task in the Tweets subset.

Model	F1	Acc.	Cc-Acc.
RobBERT	0.347	0.539	0.692
Multi-task (0.25)	0.397	0.509	0.669
Multi-task (0.5)	0.373	0.491	0.663
Multi-task (0.75)	0.372	0.482	0.655
Meta-learner	0.420	0.554	0.710
Pivot	0.281	0.426	0.651

Table 5. Pearson's r for the different models on the VAD regression task in the Tweets subset.

Model	r
RobBERT	0.635
Multi-task (0.75)	0.528
Multi-task (0.5)	0.445
Multi-task (0.25)	0.436
Meta-learner	0.638

Table 6. Macro F1, accuracy and cost-corrected accuracy for the different models on the classification task in the Captions subset.

Model	F1	Acc.	Cc-Acc.
RobBERT	0.372	0.478	0.654
Multi-task (0.25)	0.402	0.511	0.674
Multi-task (0.5)	0.408	0.504	0.663
Multi-task (0.75)	0.401	0.473	0.645
Meta-learner	0.407	0.516	0.678
Pivot	0.275	0.429	0.605

Table 7. Pearson's r for the different models on the VAD regression task in the Captions subset.

Model	r
RobBERT	0.641
Multi-task (0.75)	0.551
Multi-task (0.5)	0.540
Multi-task (0.25)	0.520
Meta-learner	0.643

The results of the base models are rather similar in both domains. As also observed in De Bruyne et al. [13], the performance is notably low for categories, especially regarding macro F1-score (only 0.347 for Tweets and 0.372 for Captions). Note that we are dealing with imbalanced datasets, which explains the discrepancy between macro F1 and accuracy (instances per category in Tweets subcorpus: $n_{\text{anger}} = 188$, $n_{\text{fear}} = 51$, $n_{\text{joy}} = 400$, $n_{\text{love}} = 44$, $n_{\text{sadness}} = 98$, $n_{\text{neutral}} = 219$; Captions subcorpus: $n_{\text{anger}} = 198$, $n_{\text{fear}} = 96$, $n_{\text{joy}} = 340$, $n_{\text{love}} = 45$, $n_{\text{sadness}} = 186$, $n_{\text{neutral}} = 135$). Scores for dimensions seem more promising, although results are hard to compare as we are dealing with different metrics ($r = 0.635$ for Tweets and 0.641 for Captions).

When we look at multi-framework settings (multi-task and metalearner), we see that performance goes up for the categories (from 0.347 to 0.420 in the meta-learning setting for Tweets and from 0.372 to 0.407 for Captions), while it drops or stays constant for the dimensions (from 0.635 to 0.638 and from 0.641 to 0.643 for the meta-learner in Tweets and Captions, respectively). This observation confirms that categories benefit more from the additional information of dimensions than in the opposite direction and corroborates the assumption that the VAD model is more robust than the classification model.

The boost in performance for categories is especially clear for the meta-learner setting, where scores improve for all evaluation metrics in both domains (increase of no less than 7% macro F1 and around 2% (cost-corrected) accuracy for Tweets and around 3% in all metrics for Captions).

For the multi-task approach, only macro F1 increased for categories, while for Captions, (cost-corrected) accuracy also went up in two out of three settings. When taking all metrics into account, the largest increase was found in the setting where VAD had the largest weight (noted in Tables 4 and 6 as Multi-task (0.25)).

For the pivot method, the primary objective was not to outperform the base model, but to be on par with it. However, looking at the performance, we observe a steep drop in performance for all metrics (e.g., for Tweets accuracy and Captions F1 the decrease is almost 10%). The loss in cost-corrected accuracy is smaller. Error analysis will need to clarify whether predictions made in the pivot approach are useful (see Section 5). However, based on these results, it does not seem that the pivot method is an effective approach to predict emotion categories.

5. Discussion

The results in Section 4 suggest that VAD dimensions can help in predicting emotional categories, as the VAD regression model seems more robust than the classification model. However, the pivot method did not seem an effective approach to predict emotion categories. In this section, we will take a look at the correlation between categories and VAD dimensions as annotated in our dataset and perform an error analysis on the predictions of the pivot method. Finally, we give some suggestions for future research directions.

5.1. Correlation between Categories and Dimensions

The point biserial correlation coefficient is used to measure correlation between a continuous and a binary variable. This allows us to assess the correlation between each emotion category (either 0 or 1, so the binary variable) and each one of the VAD dimensions (continuous). The results are shown in Table 8 (Tweets) and Table 9 (Captions).

Table 8. Point biserial correlation coefficient between VAD values and categories in the Tweets subset. * indicates that $p < 0.05$.

	V	A	D
Neutral	0.05	−0.29 *	−0.05
Anger	−0.44 *	0.08 *	0.18 *
Fear	−0.16 *	0.00	−0.20 *
Joy	0.56 *	0.20 *	0.25 *
Love	0.20 *	0.06	0.02
Sadness	−0.44 *	−0.06	−0.46 *

Table 9. Point biserial correlation coefficient between VAD values and categories in the Captions subset. * indicates that $p < 0.05$.

	V	A	D
Neutral	0.03	−0.34 *	0.08 *
Anger	−0.47 *	0.34 *	0.03
Fear	−0.11 *	0.04	−0.31 *
Joy	0.67 *	0.09 *	0.42 *
Love	0.21 *	−0.06	0.13 *
Sadness	−0.39 *	−0.16 *	−0.45 *

In both domains, *anger* and *sadness* show a high negative correlation with *valence* (Tweets subset: $r = -0.44$ and $r = -0.44$, respectively; Captions subset: $r = -0.47$ and $r = -0.39$), while *joy* shows a high positive correlation with this dimension ($r = 0.56$ for Tweets and $r = 0.67$ for Captions). For *fear* and *love*, the correlation is less obvious (Tweets: $r = -0.16$ and $r = 0.20$; Captions: $r = -0.11$ and $r = 0.21$).

Arousal is (weakly) positively correlated with *anger* and *joy* (Tweets: $r = 0.08$ and $r = 0.20$, respectively; Captions: $r = 0.34$ and $r = 0.09$). *Sadness* has a negative correlation with this dimension in Captions ($r = -0.16$). Strikingly, *neutral* has a notable negative correlation with *arousal* ($r = -0.29$ in Tweets and $r = -0.34$ in Captions). This goes a bit against our assumption that the neutral state is the center of the VAD space, although it is not completely counter-intuitive that neutral sentences were judged as having low arousal instead of medium arousal.

Contrary to what some studies claim [36], the *dominance* dimension seems more correlated with emotion categories than *arousal*. Especially with *sadness*, with which *dominance* is negatively correlated, the correlation is rather high ($r = -0.46$ in Tweets and $r = -0.45$ in Captions). In the Captions subset, *fear* and *joy* are rather highly correlated with *dominance* as well ($r = -0.31$ and $r = 0.42$, respectively).

The dimensional and categorical annotations in our dataset are thus correlated, but not for each dimension-category pair and certainly not always to a great extent. These observations do seem to suggest that a mapping could be learned. Indeed, various studies have already successfully accomplished this [19–21]. However, our goal is not to *learn* a mapping, because then there would still be a need for annotations in the target label set. Instead, a mapping should be achieved without relying on any categorical annotation. The correlations shown in Tables 8 and 9 thus seem too low to directly map VAD predictions to categories through a rule-based approach, as was proven in the results of the presented pivot method.

For comparison, we did try to learn a simple mapping using an SVM. This is a similar approach as the one depicted in Figure 3, but now only the VAD predictions are used as input for the SVM classifier. Results of this learned mapping are shown in Table 10. Especially for the Tweets subset, results for the learned mapping are on par with the base model, suggesting that a pivot method based on a learned mapping could actually be operative.

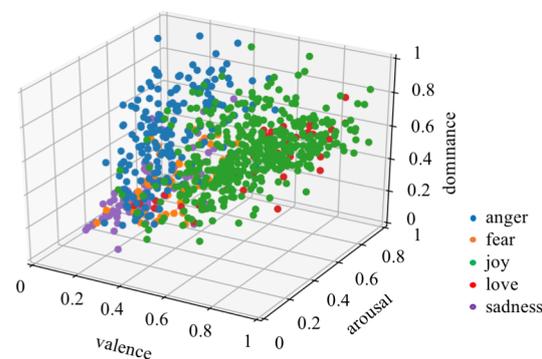
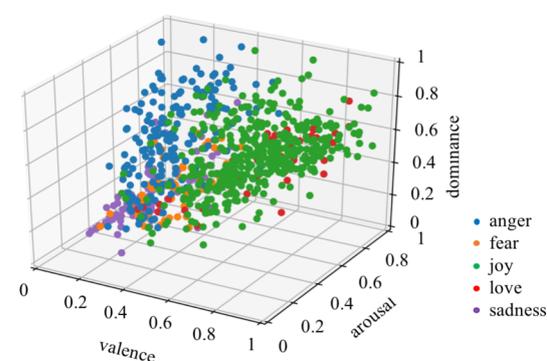
Table 10. Macro F1, accuracy and cost-corrected accuracy for the learned mapping from VAD to categories in the Tweets and Captions subset.

Model	Tweets			Captions		
	F1	Acc.	Cc-Acc.	F1	Acc.	Cc-Acc.
RobBERT	0.347	0.539	0.692	0.372	0.478	0.654
Learned mapping	0.345	0.532	0.697	0.271	0.457	0.591

Apart from looking at correlation coefficients, we also try to visualise the relation between categories and dimensions in our data. We do this by plotting each annotated instance in the three-dimensional space according to its dimensional annotation, while at the same time visualising its categorical annotation through colours.

Figures 5 and 6 visualise the distribution of data instances in the VAD space according to their dimensional and categorical annotations. On the *valence* axis, we clearly see a distinction between the *anger* (blue) and *joy* (green) cloud. In the negative *valence* area, *anger* is more or less separated from *sadness* and *fear* on the *dominance* axis, although *sadness* and *fear* seem to overlap rather strongly. In addition, *joy* and *love* show a notable overlap.

Average vectors per emotion category are shown in Figures 7 and 8. It is striking that these figures, although they are based on annotated real-life data (tweets and captions), are very similar to the mapping of individual emotion terms as defined by Mehrabian [12] (Figure 1), although the categories with higher *valence* or *dominance* are shifted a bit more to the neutral point of the space. Again, it is clear that *joy* and *love* are very close to each other, while the negative emotions (especially *anger* with respect to *fear* and *sadness*) are better separated.

**Figure 5.** Distribution of instances from the Tweets subset in the VAD space, visualised according to emotion category.**Figure 6.** Distribution of instances from the Captions subset in the VAD space, visualised according to emotion category.

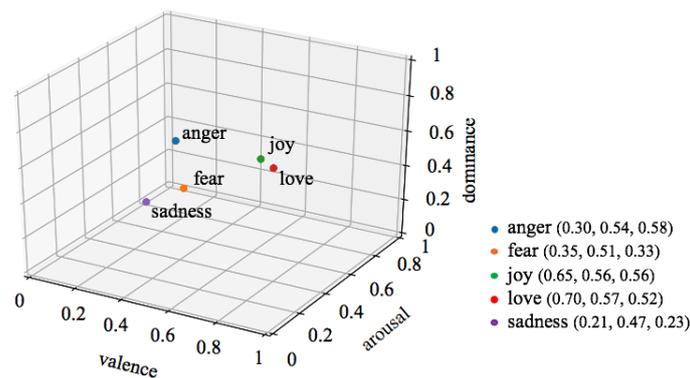


Figure 7. Average VAD vector of instances from the Tweets subset, visualised according to emotion category.

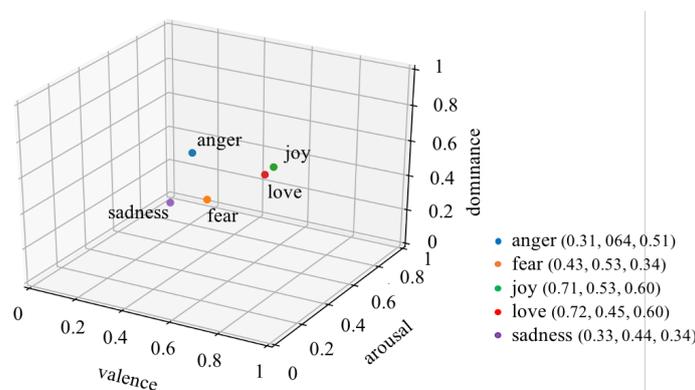


Figure 8. Average VAD vector of instances from the Captions subset, visualised according to emotion category.

Although the average VAD per category values corresponds well to the definitions of Mehrabian [12], which are used in our mapping rule, the individual data points are very much spread out over the VAD space. This results in quite some overlap between the classes. Moreover, many (predicted) data points within a class will actually be closer to the center of the VAD space than it is to the average of its class. However, this is somewhat accounted for in our mapping rule by first checking conditions and only calculating cosine distance when no match is found (see Table 3). Nevertheless, inferring emotion categories purely based on VAD predictions does not seem efficient.

5.2. Error Analysis

In order to get some more insights into the decisions of our proposed models, we perform an error analysis on the classification predictions. We show the confusion matrices of the base model, the best performing multi-framework model (which is the meta-learner) and the pivot model. Then, we randomly select a number of instances and discuss their predictions.

Confusion matrices for Tweets are shown in Figures 9–11, and those of the Captions subset are shown in Figures 12–14. Although the base model's accuracy was higher for the Tweets subset than for Captions, the confusion matrices show that there are less misclassifications per class in Captions, which corresponds to its overall higher macro F1 score (0.372 compared to 0.347). Overall, the classifiers perform poorly on the smaller classes (*fear* and *love*).

For both subsets, the diagonal in the meta-learner's confusion matrix is more pronounced, which indicates more true positives. The most notable improvement is for *fear*. Besides *fear*, *love* and *sadness* are the categories that benefit most from the meta-learning

model. There is an increase of respectively 17%, 9% and 13% F1-score in the Tweets subset and one of 8%, 4% and 6% in Captions.

The pivot method clearly falls short. In the Tweets subset, only the predictions for *joy* and *sadness* are acceptable, while *anger* and *fear* get mixed up with *sadness*. In the Captions subset, the pivot method fails to make good predictions for all negative emotions.

		PREDICTED					
		😊	😡	😭	😄	😘	😱
T	😊	86	15	0	113	0	5
	😡	39	97	0	39	0	13
R	😭	19	8	0	14	0	10
	😄	54	19	0	324	1	2
E	😘	4	1	0	36	2	1
	😱	16	28	1	23	0	30

Figure 9. Confusion matrix base model Tweets.

		PREDICTED					
		😊	😡	😭	😄	😘	😱
T	😊	87	19	3	99	1	10
	😡	20	104	12	32	0	20
R	😭	14	7	7	12	0	11
	😄	54	26	3	301	8	8
E	😘	4	1	0	33	5	1
	😱	10	20	6	10	2	50

Figure 10. Confusion matrix meta-learner Tweets.

		PREDICTED					
		😊	😡	😭	😄	😘	😱
T	😊	30	6	9	127	3	44
	😡	25	48	30	27	0	58
R	😭	4	5	5	6	1	30
	😄	30	16	16	287	23	28
E	😘	4	1	3	31	3	2
	😱	5	4	28	7	1	53

Figure 11. Confusion matrix pivot model Tweets.

		PREDICTED					
		😊	😡	😭	😄	😘	😱
T	😊	64	17	7	20	0	27
	😡	26	89	15	26	0	42
R	😭	15	10	23	11	0	37
	😄	68	27	25	186	0	34
E	😘	11	5	2	17	0	10
	😱	26	20	14	10	0	116

Figure 12. Confusion matrix base model Captions.

		PREDICTED					
		😊	😡	😭	😄	😘	😱
T	😊	60	10	12	35	0	18
	😡	25	71	27	47	1	27
R	😭	11	6	36	18	1	24
	😄	44	14	23	239	5	15
E	😘	5	2	4	31	1	2
	😱	17	13	21	26	0	109

Figure 13. Confusion matrix meta-learner Captions.

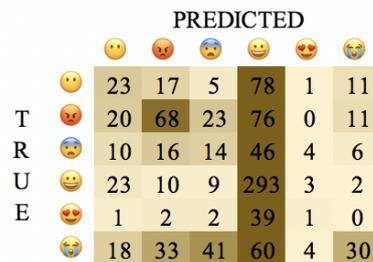


Figure 14. Confusion matrix pivot model Captions.

To get more insights into the misclassifications, ten instances (five from the Tweets subcorpus and five from Captions) were randomly selected for further analysis. These are shown in Table 11 (an English translation of the instances is given in Appendix A). In all given instances (except instance 2), the base model gave a wrong prediction, while the meta-learner outputted the correct class. In particular, the first example is interesting, as this instance contains irony. At first glance, the sunglasses emoji and the words “een politicus liegt nooit” (*politicians never lie*) seem to express joy, but context makes us understand that this is in fact an angry message. Probably, the valence information present in the VAD predictions is the reason why the polarity was flipped in the meta-learner prediction. Note that the output of the pivot method is a negative emotion as well, albeit sadness.

Table 11. Selection of instances and their gold label and predictions by the base model (RobBERT), the meta-learner and the pivot method.

	Instance	Gold	RobBERT	Meta-Learner	Pivot	VAD Gold	VAD Predicted
1	@TvdVen natuurlijk niet waar 😎 een politicus liegt nooit	anger	joy	anger	sadness	0.32, 0.29, 0.76	0.42, 0.26, 0.84
2	Da sta mij geen beetje aan eh.	anger	anger	anger	sadness	0.35, 0.41, .48	0.16, 0.68, 0.53
3	Kheb echt nog nooit zo veel zenuwen gehad. Echt nog nooit. Das één woord dak da mee kan omschrijven, das doodgaan.	fear	sadness	fear	anger	0.48, 0.67, 0.61	0.06, 0.79, 0.34
4	Het is toch eerst oefenen en dan voor het echie? Ohh 😎 deze keer niet.#afgang #bulned #itaned	fear	neutral	fear	anger	0.45, 0.59, 0.60	0.28, 0.56, 0.28
5	Ma wij gaan dus wel effectief aan zo’n machien moeten zitten eh, voor alle duidelijkheid.	fear	anger	fear	joy	0.51, 0.50, 0.65	0.55, 0.50, 0.58
6	Ik kom heel goed overeen met mijn papa. Die steunt mij ook altijd en is er altijd. En als ik mij slecht voel, kan die altijd opbellen en dan... dan geeft die mij weer zo ne peptalk en euh... Ja das mijn nummer één fan om het zo te zeggen. Dus das keileuk, ja.	love	anger	love	joy	0.79, 0.56, 0.81	0.81, 0.60, 0.70
7	Kijk nou! Wtf, deze jongen heeft al zó veel bereikt! En is vanaf vandaag pas 18 🤩 echt een voorbeeld!DOE WAT JE LEUK VINDT! ❤️ en u see, je komt er 💖 https://t.co/6AUw29DXso (accessed on 30 September 2021)	love	joy	love	love	0.64, 0.85, 0.46	0.88, 0.89, 0.64
8	Hard om te zien dat de dood van een jonge man ons doet beseffen dat we zoveel meer van het leven moeten genieten 😞 RIP Lobanzo ❤️	sadness	joy	sadness	sadness	0.39, 0.34, 0.41	0.07, 0.44, 0.15
9	Het komt nu gewoon efkes allemaal heel hard binnen.	sadness	anger	sadness	fear	0.39, 0.51, 0.41	0.20, 0.30, 0.28
10	vreselijk. Zo jong 😎 sterkte voor de nabestaande. https://t.co/3NBjWIE16D (accessed on 30 September 2021)	sadness	neutral	sadness	sadness	0.20, 0.49, 0.16	0.15, 0.46, 0.23

In three cases of the shown instances, the base model predicted emotions in the wrong polarity, which were, in turn, of the correct valence (or polarity) for the meta-learner and pivot method (instance 1, 6 and 8). Indeed, although the performance of the pivot method was low regarding macro F1 and accuracy, the cost-corrected accuracy (which takes the polarity into account) was reasonably good.

What is striking is that, out of the seven examples where the pivot failed to make a correct prediction (all instances except 7, 8 and 10), four examples would have been correctly classified based on gold VAD values. This could indicate that the main problem with the pivot method is not the mapping rule, but incorrect predictions made by the VAD regression model. To investigate this, the pivot experiments were repeated, but this time using gold VAD values instead of predicted ones. As can be observed in Table 12, results are now on par with the base classification model. This suggests that our pivot method could be efficient, provided that the VAD regression model is further improved upon.

Table 12. Macro F1, accuracy and cost-corrected accuracy for the pivot method based on gold VAD values in the Tweets and Captions subset.

Model	Tweets			Captions		
	F1	Acc.	Cc-Acc.	F1	Acc.	Cc-Acc.
RobBERT	0.347	0.539	0.692	0.372	0.478	0.654
Gold pivot	0.336	0.469	0.689	0.372	0.507	0.731

5.3. Future Work

Our experiments showed that dimensional emotion representations can help in improving the performance of emotion classification models in the EmotionNL dataset. The pivot-based approach was not successful, although we found evidence that this method might be beneficial when the VAD regression model is further improved.

This leads to several suggestions for future work. First of all, we suggest that these methods are validated on other datasets and other languages than Dutch. Furthermore, we want to improve the VAD model by testing different model architectures and investigate its effect on the usability of the pivot method. The pivot method could then be investigated when used for mapping to label sets different than the one described in this paper. Finally, an interesting research direction could be to look at other modalities, e.g., facial emotion recognition (FER). A well-known problem in FER is the poor performance in real-time testing because of the bad quality of datasets [37]. It would be interesting to investigate whether dimensional representations might be of help here as well.

6. Conclusions

In emotion detection studies, researchers usually opt for either categorical or dimensional emotion frameworks. Our previous work on Dutch emotion detection showed that the classification of emotional categories is a very challenging task, but that a regression task for predicting *valence*, *arousal* and *dominance* achieves more promising results. In this paper, we have therefore investigated whether transferring information from the regression to the classification task can improve performance. Moreover, we have examined the potential of dimensional representations to be used as a pivot mechanism, which allows tailoring emotion frameworks to specific tasks and domains.

Our results reveal that dimensional representations can indeed boost the baseline emotion classification's performance, especially in a meta-learning setting. Moreover, while categories do benefit from the additional VAD information, the opposite does not hold, which further underlines the assumption that the VAD model is more robust than the classification model.

The pivot method was not able to compete with the base model and revealed a substantial drop in performance. However, further inspection revealed that the rule-based mapping itself does perform on par with the base model when gold VAD values are used. This suggests that the pivot method could be efficient, provided that the VAD regression

model is further improved. This opens up possibilities to tailor emotion frameworks to specific tasks and domains and thus broaden their application scope.

Author Contributions: Conceptualization, L.D.B., O.D.C. and V.H.; methodology, L.D.B., O.D.C. and V.H.; software, L.D.B.; validation, L.D.B., O.D.C. and V.H.; formal analysis, L.D.B.; investigation, L.D.B.; resources, L.D.B., O.D.C. and V.H.; data curation, L.D.B.; writing—original draft preparation, L.D.B.; writing—review and editing, O.D.C. and V.H.; visualization, L.D.B.; supervision, O.D.C. and V.H.; project administration, L.D.B., O.D.C. and V.H.; funding acquisition, L.D.B., O.D.C. and V.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Research Foundation–Flanders under a Strategic Basic Research fellowship with Grant No. 3S004019 (<https://www.researchportal.be/nl/project/transfer-learning-voor-automatische-erotiedetectie-nederlandstalige-teksten>) (accessed on 30 September 2021).

Data Availability Statement: The dataset will be made available for research purposes upon request at <https://lt3.ugent.be/resources/> (accessed on 30 September 2021). All source code is available at <https://github.com/LunaDeBruyne/Mixing-Matching-Emotion-Frameworks> (accessed on 30 September 2021).

Acknowledgments: We would like to thank Bram Vanroy for developing a text classification pipeline on which the code for these experiments is based. The base code can be found at <https://github.com/BramVanroy/lt3-2019-transformer-trainer> (accessed on 30 September 2021).

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Appendix A

Table A1. English translation of the selection of instances from Table 11.

Instance
1 @TvdVen of course not trye 🤔 a politician never lies
2 I don't like that at all.
3 I've never been so nervous. Really never. There is one word I can describe it with, which is dying.
4 It is first practice and then for real, right? Ohh 🤔 not this time.#fail #bulned #itaned
5 However, we will really have to sit at such a machine, just to be clear.
6 I get along really well with my dad. He always supports me and he is always there. In addition, when I'm feeling down, I can always call him and... and he gives me a peptalk and ehm... Yes he is my number one fan so to speak. Thus that is super nice.
7 Look!Wtf, this guy has achieved already so much!En he only turned 18 today 🤔 a real example!DO WHAT YOU LIKE! ❤️ and u see,you'll get there 💖 https://t.co/6AUw29DXso (accessed on 30 September 2021)
8 Tough to see that the death of a young man makes us realise we should enjoy life so much more 🤔 RIP Lobanzo ❤️
9 It's just really starting to hit me now.
10 terrible. Thus, young 🤔 condolences to the grieving family. https://t.co/3NBjWIE16D (accessed on 30 September 2021)

References

- Liu, B. Sentiment analysis and opinion mining. *Synth. Lect. Hum. Lang. Technol.* **2012**, *5*, 1–167. [CrossRef]
- El Marrakchi, M.; Bellafkih, M.; Bensaid, H. Towards reputation measurement in online social networks. In Proceedings of the 2015 IEEE Intelligent Systems and Computer Vision (ISCV), Fez, Morocco, 25–26 March 2015; pp. 1–8.
- Pontiki, M.; Galanis, D.; Pavlopoulos, J.; Papageorgiou, H.; Androutsopoulos, I.; Manandhar, S. SemEval-2014 Task 4: Aspect Based Sentiment Analysis. In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland, 23–24 August 2014; pp. 27–35.
- Mohammad, S.M. Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In *Emotion Measurement*; Woodhead Publishing: Cambridge, UK, 2016; pp. 201–237.
- Ekman, P. An Argument for Basic Emotions. *Cogn. Emot.* **1992**, *6*, 169–200. [CrossRef]
- Plutchik, R. A General Psychoevolutionary Theory of Emotion. In *Theories of Emotion*; Plutchik, R., Kellerman, H., Eds.; Academic Press: New York, NY, USA, 1980; pp. 3–33.
- Buechel, S.; Hahn, U. Emotion Analysis as a Regression Problem—Dimensional Models and Their Implications on Emotion Representation and Metrical Evaluation. In Proceedings of the 22nd European Conference on Artificial Intelligence—ECAI 2016, The Hague, The Netherlands, 29 August–2 September 2016; pp. 1114–1122.

8. Mohammad, S.; Kiritchenko, S. Understanding Emotions: A Dataset of Tweets to Study Interactions between Affect Categories. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 7–12 May 2018; pp. 198–209.
9. De Bruyne, L.; De Clercq, O.; Hoste, V. An Emotional Mess! Deciding on a Framework for Building a Dutch Emotion-Annotated Corpus. In Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, 11–16 May 2020; pp. 1643–1651.
10. Jin, Y.; Liu, B.F.; Anagondahalli, D.; Austin, L. Scale development for measuring publics' emotions in organizational crises. *Public Relat. Rev.* **2014**, *40*, 509–518. [[CrossRef](#)]
11. Hoste, V.; Van Hee, C.; Poels, K. Towards a framework for the automatic detection of crisis emotions on social media: A corpus analysis of the tweets posted after the crash of germanwings flight 9525. In Proceedings of the HUSO 2016, the Second International Conference on Human and Social Analytics, Barcelona, Spain, 13–17 November 2016; pp. 29–32.
12. Mehrabian, A.; Russell, J.A. *An Approach to Environmental Psychology*; MIT Press: Cambridge, MA, USA, 1974.
13. De Bruyne, L.; De Clercq, O.; Hoste, V. Prospects for Dutch Emotion Detection: Insights from the new EmotioNL Dataset. *Comput. Linguist. Neth. J.* **2021**, submitted.
14. Augenstein, I.; Ruder, S.; Søgaard, A. Multi-Task Learning of Pairwise Sequence Classification Tasks over Disparate Label Spaces. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), New Orleans, LA, USA, 1–6 June 2018; pp. 1896–1906.
15. Chauhan, D.S.; S R, D.; Ekbal, A.; Bhattacharyya, P. 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, Online, 5–10 July 2020; pp. 4351–4360.
16. Akhtar, S.; Ghosal, D.; Ekbal, A.; Bhattacharyya, P.; Kurohashi, S. All-in-One: Emotion, Sentiment and Intensity Prediction using a Multi-task Ensemble Framework. *IEEE Trans. Affect. Comput.* **2019**, *1*. [[CrossRef](#)]
17. Akhtar, M.S.; Chauhan, D.; Ghosal, D.; Poria, S.; Ekbal, A.; Bhattacharyya, P. Multi-task Learning for Multi-modal Emotion Recognition and Sentiment Analysis. 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), Minneapolis, MN, USA, 3–5 June 2019; pp. 370–379.
18. Xu, P.; Madotto, A.; Wu, C.S.; Park, J.H.; Fung, P. Emo2Vec: Learning Generalized Emotion Representation by Multi-task Training. In Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Brussels, Belgium, 31 October 2018; pp. 292–298.
19. Stevenson, R.A.; Mikels, J.A.; James, T.W. Characterization of the Affective Norms for English Words by discrete emotional categories. *Behav. Res. Methods* **2007**, *39*, 1020–1024. [[CrossRef](#)] [[PubMed](#)]
20. Buechel, S.; Hahn, U. A Flexible Mapping Scheme for Discrete and Dimensional Emotion Representations. In Proceedings of the 39th Annual Meeting of the Cognitive Science Society—CogSci 2017, London, UK, 16–29 July 2017; pp. 180–185.
21. Buechel, S.; Hahn, U. Emotion Representation Mapping for Automatic Lexicon Construction (Mostly) Performs on Human Level. In Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, NM, USA, 20–26 August 2018; pp. 2892–2904.
22. Calvo, R.A.; Mac Kim, S. Emotions in text: Dimensional and categorical models. *Comput. Intell.* **2013**, *29*, 527–543. [[CrossRef](#)]
23. Bradley, M.M.; Lang, P.J. *Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings*; Technical Report; University of Florida: Gainesville, FL, USA, 1999.
24. Emerson, G.; Declerck, T. SentiMerge: Combining Sentiment Lexicons in a Bayesian Framework. In Proceedings of the Workshop on Lexical and Grammatical Resources for Language Processing, Dublin, Ireland, 24 August 2014; pp. 30–38.
25. Hoyle, A.M.; Wolf-Sonkin, L.; Wallach, H.; Cotterell, R.; Augenstein, I. Combining Sentiment Lexica with a Multi-View Variational Autoencoder. 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), Minneapolis, MN, USA, 3–5 June 2019; pp. 635–640.
26. De Bruyne, L.; Atanasova, P.; Augenstein, I. Joint emotion label space modeling for affect lexica. *Comput. Speech Lang.* **2022**, *71*, 101257. [[CrossRef](#)]
27. Bostan, L.A.M.; Klinger, R. An Analysis of Annotated Corpora for Emotion Classification in Text. In Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, NM, USA, 21–25 August 2018; pp. 2104–2119.
28. Delobelle, P.; Winters, T.; Berendt, B. RobBERT: A dutch RoBERTa-based language model. *arXiv* **2020**, arXiv:2001.06286.
29. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
30. Suárez, P.J.O.; Sagot, B.; Romary, L. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In Proceedings of the 7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7), Cardiff, UK, 22 July 2019.
31. de Vries, W.; van Cranenburgh, A.; Bisazza, A.; Caselli, T.; van Noord, G.; Nissim, M. Bertje: A dutch bert model. *arXiv* **2019**, arXiv:1912.09582.
32. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv* **2019**, arXiv:1910.03771.
33. Loshchilov, I.; Hutter, F. Fixing weight decay regularization. *arXiv* **2017**, arXiv:1711.05101.

-
34. Hendrycks, D.; Gimpel, K. Gaussian error linear units (gelus). *arXiv* **2016**, arXiv:1606.08415.
 35. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
 36. Russell, J.A. A circumplex model of affect. *J. Personal. Soc. Psychol.* **1980**, *39*, 1161. [[CrossRef](#)]
 37. Kim, J.H.; Poulouse, A.; Han, D.S. The Extensive Usage of the Facial Image Threshing Machine for Facial Emotion Recognition Performance. *Sensors* **2021**, *21*, 2026. [[CrossRef](#)] [[PubMed](#)]