Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 1643–1651
Marseille, 11–16 May 2020

An Emotional Mess! Deciding on a Framework for Building a Dutch Emotion-Annotated Corpus

Luna De Bruyne, Orphée De Clercq, Veronique Hoste
LT^3, Language and Translation Technology Team, Ghent University
Groot-Brittanniëlaan 45, 9000 Ghent, Belgium
{luna.debruyne, orphee.declercq, veronique.hoste}@ugent.be

Abstract

Seeing the myriad of existing emotion models, with the categorical versus dimensional opposition the most important dividing line, building an emotion-annotated corpus requires some well thought-out strategies concerning framework choice. In our work on automatic emotion detection in Dutch texts, we investigate this problem by means of two case studies. We find that the labels *joy*, *love*, *anger*, *sadness* and *fear* are well-suited to annotate texts coming from various domains and topics, but that the connotation of the labels strongly depends on the origin of the texts. Moreover, it seems that information is lost when an emotional state is forcedly classified in a limited set of categories, indicating that a bi-representational format is desirable when creating an emotion corpus.

Keywords: NLP, emotion detection, emotion annotation

1. Introduction

When dealing with the task of automatically detecting emotions in texts – a well-studied topic in the field of natural language processing or NLP (Mohammad et al., 2018; Chatterjee et al., 2019) – the first bottleneck is data acquisition. Not only is there the need to collect a considerable amount of data, one also needs to decide on an appropriate framework to annotate these data instances in order to build an emotion corpus. Seeing the plethora of existing emotion frameworks, this decision is not trivial.

On the one hand, emotions can be represented as categories, typically by using a set of basic emotions. The frameworks of Ekman (1992) (with the basic emotions *anger*, *disgust*, *fear*, *joy*, *sadness* and *surprise*) and Plutchik (1980) (in which *trust* and *anticipation* are added) are the most popular ones, but many other theorists provided sets of basic emotions, counting up to 14 emotion categories, e.g. the emotion theory of Roseman (1984).

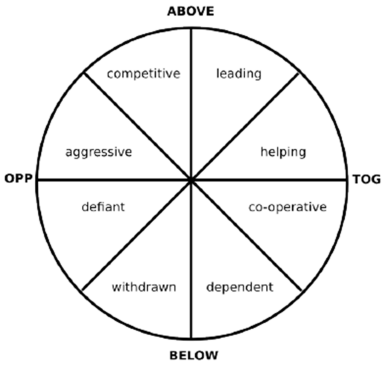
Dimensional models, on the other hand, represent emotions as vectors in a multidimensional space. Mehrabian and Russell (1974) claimed that this emotional space is defined by the axes *valence* (unhappiness-happiness), *arousal* (calmness-excitement) and *dominance* (submission-dominance), so that every emotional state can be described as a combination of values on these VAD-axes. However, in later work, Russell (1980) argued that only the dimensions *valence* and *arousal* are necessary for describing any emotional state, whereas Fontaine et al. (2007) claim a fourth dimension, *unpredictability*, is needed.

Building an emotion-annotated corpus thus requires some well thought-out strategies concerning annotation method and framework choice. However, in most emotion datasets, the motives on which a particular emotion framework is selected are unclear or the choice even seems arbitrary. In almost all studies, preference is given to categorical approaches for building datasets, but only very rarely a motivation, let alone an experimentally grounded one, is given. Moreover, in psychology, where these frameworks originate from, not a single study deals with fully-fledged textual data. The theory of Ekman (1992) is based on visual expressions, and Plutchik (1980) built his wheel of emotions on psychoevolutionary and behavioral observations. The works of Shaver et al. (1987) (in which a similarity-sorting task and cluster analysis of a large number of emotion terms led to an extensive emotion taxonomy) and Mehrabian and Russell (1974) (who performed rating studies of emotion-eliciting situations, but also of emotional words on the VAD-dimensions) do work with textual data, but stick to word-level experiments. This raises questions about the validity of these frameworks for research in NLP. How should we then, in this jumble of emotion models, decide on a framework, so that our corpus is experimentally grounded, usable and reliable?

To tackle this problem, De Bruyne et al. (2019) established an emotion framework based on a cluster analysis on real-life data (in their case: Dutch Twitter messages), justifying their framework both theoretically and practically. This resulted in an experimentally grounded label set consisting of the five emotions *joy*, *love*, *anger*, *nervousness* and *sadness*. However, the study only focused on categorical models.

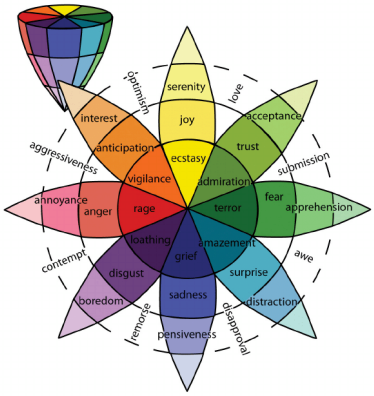
In this work, we will pursue the work by De Bruyne et al. (2019) and investigate label sets by means of two case studies, using real-life data. Study 1 addresses categorical annotations and the need for an experimentally grounded label set in two different domains, namely Dutch twitter messages and subtitles from reality TV shows, in order to investigate the importance of domain and topic on the label set. Study 2 explores dimensional annotations in the Twitter domain. We investigate the robustness of the VAD-model on real-life data, and examine how emotional categories relate to dimensions. This way, we can validate the use of both dimensional and categorical models for labeling 'texts in the wild', going beyond word-level experiments.

We find that the labels found by De Bruyne et al. (2019) (*joy*, *love*, *anger*, *sadness* and *nervousness* – or more generally: *fear*), are well-suited to annotate texts coming from

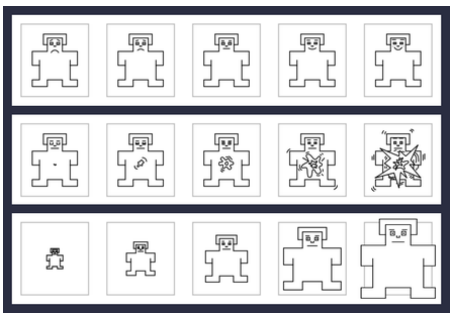


(a) Leary's rose, divided into the octants *leading*, *helping*, *co-operative*, *dependent*, *withdrawn*, *defiant*, *aggressive*, *competitive* (Leary, 1957).

(b) Ekman's basic six: *anger*, *fear*, *disgust*, *surprise*, *joy*, *sadness* (Ekman, 1992).



(c) Plutchik's wheel of emotions with basic emotions *joy*, *trust*, *fear*, *surprise*, *sadness*, *disgust*, *anger*, *anticipation* (Plutchik, 1980).



(d) The self-assessment manikin, a tool for annotating the emotional dimensions *valence*, *arousal* and *dominance* (Bradley and Lang, 1994).

Figure 1: Examples of emotion frameworks.

various domains, but that the connotation of the labels strongly depends on the origin of the texts. Moreover, it seems that the categories *joy* and *love* cannot express all nuances of positive emotional states, indicating that information is lost when emotional states are forcedly classified in a limited set of categories.

This paper begins by discussing related work on frameworks for building emotion datasets in Section 2. It will then go on to the description of the experiments and its results (Section 3), with Study 1 on categorical label sets in Section 3.1 and Study 2 on dimensional annotations in Section 3.2. We end this paper with some concluding thoughts and suggestions in Section 4.

2. Related Work

2.1. Dutch emotion corpora

The availability of Dutch emotion-annotated corpora is extremely restricted. The `deLearyous` dataset (Vaassen and Daelemans, 2011) is the only publicly available Dutch emotion dataset known to us. The dataset consists of 740 Dutch sentences from conversations, annotated according to Leary's Rose or the Interpersonal Circumplex (Leary, 1957). Leary's Rose is a circumplex model defined by the two axes *opposite-together* (willingness to cooperate with listener) and *above-below* (how dominant or submissive the speaker is towards the listener), resulting in eight octants (*leading*, *helping*, *co-operative*, *dependant*, *withdrawn*, *defiant*, *aggressive*, *competitive*), as shown in Figure 1a.

The choice for Leary's Rose as framework is remarkable, not only because it is the first study that uses this framework in the scope of emotion detection, but especially because it was developed for structuring interpersonal behavior in personality theory, rather than emotion theory. Moreover, in his doctoral thesis, Vaassen (2014) concludes that the performance of his machine learning system for classifying sentences into the quadrants or octants of Leary's Rose is too low to be used in practice, and attributes this to shortcomings in gold-standard data: the data are too sparse, and the inter-annotator agreement is low (Fleiss Kappa of 0.37 for quadrants and 0.29 for octants). He suggests collecting larger datasets with higher agreement of emotion annotations for future work and encourages to work within a cross-domain framework to expand datasets.

2.2. Categorical frameworks in emotion corpora

When looking at other languages, we see that a substantial number of emotion corpora have been created by NLP researchers, but almost all of them focus on the English language. Categorical frameworks are dominant, and both Ekman's (Figure 1b) and Plutchik's (Figure 1c) set of basic emotions are popular. Even more often, variations of these

Dataset	Reference	Framework
Affect in Tweets	(Mohammad et al., 2018)	Plutchik + *optimism*, *pessimism*, *love*
Affective Text	(Strapparava and Mihalcea, 2007)	Ekman
Blogs	(Aman and Szpakowicz, 2007)	Ekman + no emotion, mixed emotion
Daily-Dialog	(Li et al., 2017)	Ekman + no emotion
Electoral Tweets	(Mohammad et al., 2015)	*Acceptance*, *admiration*, *amazement*, *anger*, *anticipation*, *calmness*, *disappointment*, *disgust*, *dislike*, *fear*, *hate*, *indifference*, *joy*, *like*, *sadness*, *surprise*, *trust*, *uncertainty*, *vigilance*
EmoBank	(Buechel and Hahn, 2017)	*Valence*, *arousal*, *dominance* + subset with additional Ekman annotations
EmoInt	(Mohammad and Bravo-Marquez, 2017)	*Anger*, *fear*, *joy*, *sadness*
Emotion in Text	(Figure Eight/CrowdFlower, 2016)	*Anger*, *disgust*, *fear, joy*, *sadness*, *surprise*, *enthusiasm*, *fun*, *hate*, *love*, *boredom*, *relief*, *empty*, *neutral*
Emotion-Stimulus	(Ghazi et al., 2015)	Ekman + *shame*
Facebook-VA	(Preoţiuc-Pietro et al., 2016)	*Valence*, *arousal*
ISEAR	(Scherer and Wallbott, 1994)	Ekman + *guilt*, *shame*
SSEC	(Schuff et al., 2017)	Plutchik
Tales	(Alm et al., 2005)	Ekman with *positive surprise* and *negative surprise* + neutral
TEC	(Mohammad, 2012)	Ekman

Table 1: Most commonly used English emotion datasets and their frameworks.

frameworks are used (see Table 1 for an overview of the most used English emotion datasets and their frameworks). In an attempt to combat the plethora of emotion frameworks used in emotion detection studies, Schröder et al. (2006) expressed the need for a standardized model for annotating emotions. This resulted in the creation of the Emotion Annotation and Representation Language (EARL), but its construction was not data-driven nor experimentally grounded. Moreover, rather than having a universal and standardized framework, one could give preference to a well-motivated model, adjusted to the domain and task at hand. Pestian et al. (2012), for example, organised a shared task on emotion detection in suicide notes and employed a set of 15 emotions which might be indicative of suicidal behavior. However, tailoring the framework to the specific task or domain of the data only happens in rare cases, and in most studies, the motives for choosing a particular emotion framework are unclear (De Bruyne et al., 2019).

2.3. Dimensional frameworks in emotion corpora

Although dimensional models are used to a lesser extent in emotion detection, some researchers recently emphasized the potential of and even need for a dimensional approach (Buechel and Hahn, 2016; Wood et al., 2018). Buechel and Hahn (2016) consider a VAD-approach superior to a categorical one due to the lack of consensus on the basic emotions set, but also because basic emotions are not equally distributed along the *valence* and *arousal* dimensions. This considering, EmoBank was created by Buechel and Hahn (2017) in a bi-representational format: 10k sentences were annotated with VAD-scores, of which a subset also has annotations for Ekman's six. The dimensional annotations were obtained using the 5-point self-assessment manikin or SAM-scale (see Figure 1d), a pictorial scale depicting the VAD-dimensions (Bradley and Lang, 1994).
Some studies have used a dimensional approach to categorical annotation, meaning that category ratings are used instead of discrete classes. Mohammad and Bravo-Marquez (2017) obtained ratings for the intensity of *anger*, *fear*, *joy* and *sadness* by using Best-Worst scaling. In this approach – which they claim to give more reliable scores than rating scales – annotators are given four items, of which they have to indicate which one is most representative for a certain emotion category (or highest on the emotional axis) and which one is not at all (lowest on the axis). This information is then converted into real-valued scores.

3. Experiments

The creation of a new Dutch emotion corpus is a necessity, as also claimed by Vaassen (2014). In order to build a corpus that is well-motivated and experimentally grounded, we explore both categorical and dimensional annotations on our data. Therefore, we perform two studies, one exploring categorical label sets obtained by a data-driven approach, and a subsequent study investigating how these categorical labels relate to emotional dimensions and how emotions described by the VAD-model behave on real-life data.

3.1. STUDY 1: Comparing categorical label sets obtained by a data-driven approach

3.1.1. Goal

In most studies on automatic emotion detection, the motives on which a particular emotion framework is selected are unclear or the choice even seems arbitrary. De Bruyne et al. (2019) combatted this problem and performed a cluster analysis of 25 emotion terms, based on the annotations of real-life data (Dutch Twitter messages). This resulted in a label set with the five emotions *joy*, *love*, *anger*, *nervousness* and *sadness*.

Claiming that the framework needs to be derived from experiments on real-life data from the same distribution as the data that is used for emotion detection, one could assume that experiments on data from other domains and even topics lead to other label sets. To verify this, we repeat the experiments done in the work of De Bruyne et al. (2019) on a new domain, namely subtitles of two Flemish reality TV shows, and compare the resulting clusters between tweets and subtitles and between different TV shows.

3.1.2. Method

Data Apart from the data used in the study of De Bruyne et al. (2019), namely Dutch Twitter messages (Tweet dataset), we collect additional data coming from a different domain: subtitles from reality TV shows (Subtitles dataset).

A total of six episodes of two Flemish reality TV shows (*Blind getrouwd* and *Bloed, zweet en luxeproblemen*, see Figure 2) were transcribed. In *Blind getrouwd*, couples that never have met before get married, based on a match made by a team of experts. In the course of six weeks they test their match, after which they can decide to stay together or divorce. *Bloed, zweet en luxeproblemen* is a docuseries in which six adolescents travel to Asia and Africa to experience how luxury products are made. For three weeks, they get immersed in the life of inequality and hard work.

(a) Still from *Blind getrouwd*.

(b) Still from *Bloed, zweet en luxeproblemen*.

Figure 2: Transcribed TV shows.

Emotion	κ	*Emotion*	κ
Anger	0,510	Lust	1
Contentment	0,461	Nervousness	0,315
Disappointment	0,188	Optimism	0,438
Disgust	0,328	Pity	0,597
Enthrallment	0,386	Pride	0,524
Enthusiasm	0,431	Rejection	0,357
Envy	nan	Relief	nan
Fear	0,254	Remorse	0,602
Frustration	0,644	Sadness	0,678
Irritation	0,423	Suffering	0,442
Joy	0,529	Surprise	0,079
Longing	0,528	Torment	0,01
Love	0,061		

Table 2: IAA scores per emotion category.

The spontaneous utterances of the participants in the shows are transcribed, using a literal transcription method. From these transcripts, 300 utterances (sentences or short sequences of sentences) were chosen, roughly screened on the presence of emotional content.

Procedure Following De Bruyne et al. (2019), we label the utterances from the TV shows with a large set of 25 emotions obtained from Shaver et al. (1987), namely *anger*, *contentment*, *disappointment*, *disgust*, *enthrallment*, *enthusiasm*, *envy*, *fear*, *frustration*, *irritation*, *joy*, *longing*, *love*, *lust*, *nervousness*, *optimism*, *pity*, *pride*, *rejection*, *relief*, *remorse*, *sadness*, *suffering*, *surprise*, *torment*. The annotators were asked to project themselves into the speaker's perspective and indicate which of the 25 emotions were (explicitly or implicitly) expressed by the speaker. There was no minimum or maximum for the number of emotions indicated as present.

The 300 utterances were labeled by a team of three experienced linguists, by splitting the dataset in batches of 100 utterances. The first annotator labeled the first and last batch, the second annotator labeled batch 2 and 3, and the third annotator only labeled the last batch. Inter-annotator agreement was calculated with Cohen's Kappa (κ) between each annotator pair on batch 3, and the mean of those two scores was taken. As shown in Table 2, IAA varied largely depending on the emotion category. A fair ($0.2 < \kappa < 0.4$) to moderate ($0.4 < \kappa < 0.6$) agreement was observed for most categories. When the emotions that were never indicated as present by at least one annotator (*envy*, *relief* and *torment*) were disregarded, we obtained an average Kappa score of 0.444 (moderate agreement). The average Kappa score between the first two annotators was 0.438; between the last two 0.475; and 0.402 between annotator 1 and 3.

The annotations for the first batch were taken by annotator 1, of the second batch by annotator 2 and the last batch by

annotator 3. Utterances for which not a single emotion was indicated as present were considered objective and were excluded from further analysis, resulting in a final Subtitles set of 293 emotional utterances, of which 151 are from *Blind getrouwd* and 142 from *Bloed, zweet en luxeproblemen*.

Considering the annotations as vectors per emotion category, we end up with 25 n-dimensional vectors, with n depending on the number of utterances taken into account. We construct a 25x25 distance matrix for the total Subtitles set and for the two subsets by measuring the Dice dissimilarity between each emotion vector pair. This is used as input for a hierarchical cluster analysis, with Ward's method (Ward Jr, 1963) as clustering algorithm.

First, we look at the newly annotated Subtitles set and do a frequency analysis. Then, a cluster analysis is executed, first on a merged set of both TV shows, then for the two shows separately. The dendrograms resulting from this hierarchical cluster analysis are compared between domains, namely Subtitles and Tweets (dendrograms from Twitter data taken from De Bruyne et al. (2019)), and between different topics, namely the different TV shows (*Blind getrouwd* en *Bloed, zweet en luxeproblemen*).

We validate the label sets obtained by the cluster analysis by mapping them in the VAD-space. The mapping is based on the ratings of the label set's terms in the Dutch VAD-lexicon of Moors et al. (2013). This lexicon was developed in the field of psychology and consists of 4,300 Dutch words, rated on *valence*, *arousal*, *dominance* and age of acquisition. The mapping will show the distribution of the new labels along the VAD-axes.

3.1.3. Results

Frequency analysis Figure 3 shows the frequencies of emotion categories indicated as present in the subsets of *Blind getrouwd* and *Bloed, zweet en luxeproblemen*.

Seeing that some emotions are underrepresented in the data, we decide to leave the categories with fewer than 10 instances in the full dataset and fewer than 5 instances in the subsets out of consideration, meaning that *envy* and *lust* are discarded in the full dataset, *envy*, *lust* and *longing* from the *Bloed, zweet en Luxeproblemen* set and *anger*, *disgust*, *envy*, *pity*, *relief* and *torment* from the *Blind getrouwd* set.

We can already observe some striking, though expected, differences between the emotion frequencies of the two TV shows. In *Blind getrouwd*, the three emotions most indicated as present are *contentment*, *joy* and *nervousness*, while the top-3 consists of *sadness*, *frustration* and *disgust* in *Bloed, zweet en luxeproblemen*. Overall, positive emotions dominate in *Blind getrouwd* and negative emotions in *Bloed, zweet en luxeproblemen*, although there are more positive ones in the latter than negatives in the former.

Cluster analysis Figure 4a shows the dendrogram of the hierarchical cluster analysis of the complete Subtitles dataset (293 utterances), without the categories *envy* and *lust*. Same as in the Tweets dendrogram (Figure 4b, obtained from De Bruyne et al. (2019)), the first separation of the tree makes the distinction between positive emotions on the left-hand side and negative emotions on the right.

For both Subtitles and Tweets, the positive emotions separate into two clusters, one with terms related to *joy* and one with terms related to *love* (although the composition differs). However, the negative side looks somewhat different. In the Tweets dendrogram, the negative emotions cluster around the categories *anger*, *sadness* and *nervousness* (related to *fear*), but in the Subtitles tree, *sadness* and *anger* are clustered together, giving space to a new cluster containing *remorse*, *disgust* and *pity*. We suspect that the latter cluster is very much influenced by the utterances from *Bloed, zweet en luxeproblemen*. The third negative cluster has *suffering*/*torment* and *fear*/*nervousness* as its cores, strangely accompanied by *surprise*. However, the many utterances in *Blind getrouwd* where the participants are nervous, shocked and surprised by the news of their blind wedding, could explain the composition of these clusters.

Seeing the influence of particular shows on the overall clustering, it makes sense to split up the clustering per TV show. Indeed, we see that *nervousness* and *surprise* are very close to each other in the *Blind getrouwd* dendrogram. What is most striking, however, is that this *fear-surprise* cluster is placed on the positive side of the tree, probably also due to this pre-wedding feeling, which is, apart from shocking and frightening, especially exciting.

For *Bloed, zweet en luxeproblemen*, the tree is more similar to the one for Tweets. *Surprise* is here clustered together with *sadness*, representing the participants' shock when being confronted with all the distress and inequality in the communities they visit. *Disgust*, which is – if it is not seen as a basic emotion on its own – most commonly merged with *anger*, is here clustered together with *sadness* as well, probably for the same reason.

We thus can conclude that the labels *joy*, *love*, *anger*, *sadness* and *nervousness* (or more generally: *fear*), are well suited to annotate texts coming from various domains, but that the composition of the clusters and the connotation of the labels strongly depend on the origin of the texts.

Mapping in VAD-space The labels *joy*, *love*, *anger*, *sadness* and *fear* have been selected by means of a data-driven approach, making them more experimentally grounded

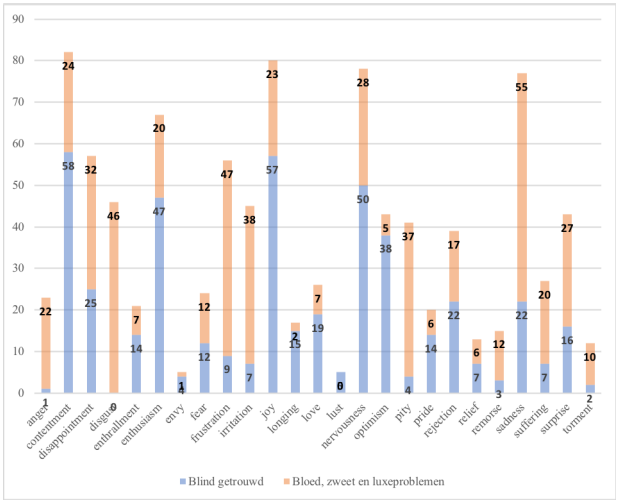
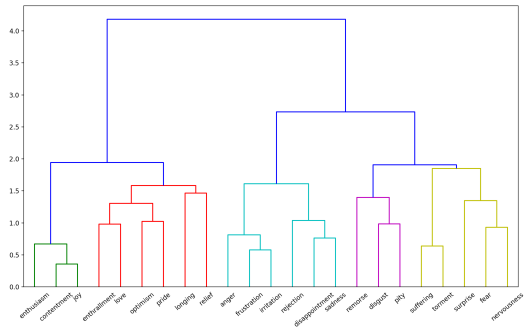
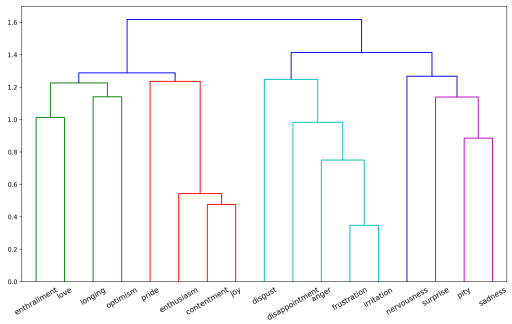


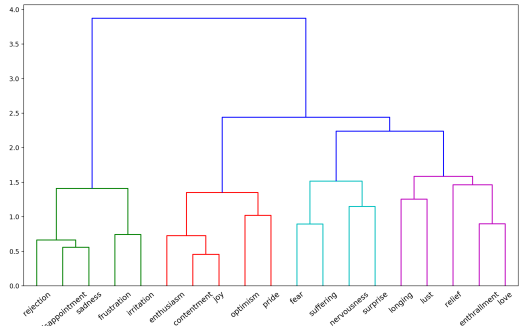
Figure 3: Frequencies of emotion categories in *Blind getrouwd* en *Bloed, zweet en luxeproblemen*.



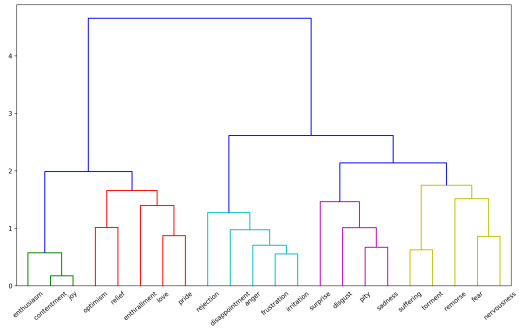
(a) Dendrogram for the Subtitles data - both TV shows.



(b) Dendrogram for the Twitter data.



(c) Dendrogram for subset *Blind getrouwd*.



(d) Dendrogram for subset *Bloed, zweet en luxeproblemen*.

Figure 4: Output of the hierarchical cluster analysis, represented by dendrograms.

than the randomly chosen label sets used in most studies on emotion detection. However, another criticism on the commonly used label sets, more specifically on Ekman's six, is that these emotion categories are not equally distributed along the axes of *valence* and *arousal* (Buechel and Hahn, 2016).

Indeed, in Ekman's six, only one category is unambiguously positive (*joy*), while *anger*, *sadness* and generally also *fear*, are negative. *Surprise*, however, can be either positive or negative. By discarding *disgust* as a separate category and adding *love*, this distribution becomes a bit more equal, at least regarding *valence*.

We show this graphically by mapping the emotion categories of Ekman into the VAD-space and we compare it with our proposed labels (Figure 5). The mapping is based on the ratings of the label set's terms in the Dutch VAD-lexicon of Moors et al. (2013). The mapping of these Dutch emotion terms is very similar to the English mapping by Mehrabian and Russell (1974) (Figure 6), with the exception that *surprise* is placed more neutrally on the *valence* axis compared to the English version (where it has a more positive score). This is also more in line with the observation that *surprise* can be both positive and negative.

Contrary to emotion categories, for which we showed their interpretation and connotation is domain-dependent, dimensional annotations seem fairly robust, at least across different languages/rating studies. Moreover, seeing the variation in interpretability of emotional categories, dimensional annotations could offer an important added value to emotion corpora, reinforcing the need for dimensional or bi-representational datasets.

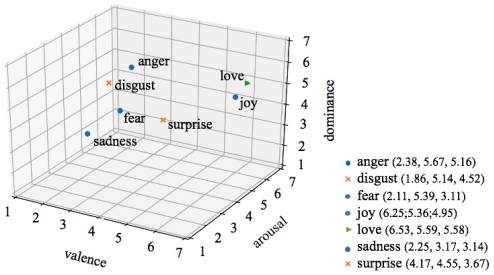


Figure 5: Mapping of Ekman's six and our proposed labels into the VAD-space. Figure based on the VAD-scores for the Dutch emotion terms in the lexicon of Moors et al. (2013).

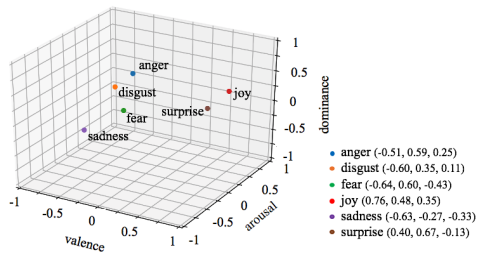


Figure 6: Mapping of Ekman's six into the VAD-space. Figure based on the scores for the English Ekman terms of Mehrabian and Russell (1974).

3.2. STUDY 2: Validating dimensional annotations

3.2.1. Goal

When using a dimensional emotion framework, the VAD-model is the most common one. According to Mehrabian and Russell (1974), every emotional state can be described using the dimensions *valence*, *arousal* and *dominance*. For illustration, they placed several emotional terms, including Ekman's basic emotions, in the 3-dimensional space (see Figure 6). Following the line of thought of the study in 3.1, we want to investigate if real-life data ('sentences in the wild'), can be mapped in a similar way in the VAD-space. Assuming that every sentence is written under a certain emotional state, sentences with a similar state should be mapped close to each other. We will link this to the emotional categories obtained in study 3.1, to evaluate a) the validity of the VAD-model (is it also robust for sentences?) and b) the usefulness of the categories obtained in study 3.1 (can we find clusters in the VAD-space that correspond to these categories?).

3.2.2. Method

Data For this study, 1000 tweets were labeled with a) the categories obtained from Study 1 (Section 3.1) and b) scores for the dimensions *valence*, *arousal* and *dominance*. The tweets were collected as described by De Bruyne et al. (2019).

Procedure For the categorical annotation, a single-label method is used, meaning that the annotators had to choose one out of the five emotions *joy*, *love*, *anger*, *fear* and *sadness*. Related emotion categories and terms from the clusters were given to the annotators, so that they had a clear comprehension of what the emotion categories consisted of. When these emotions were not uttered in the tweet, the annotators could label the tweet with either *other* or *neutral*. However, the *other* category was only very rarely indicated (15 cases) and could be omitted by replacing it with one of the five emotion categories in a revision round.

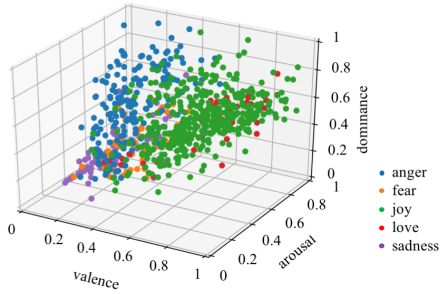
Two annotators executed this task. The dataset was split in two parts, so that both annotators labeled 500 sentences. Additionally, annotator 2 labeled 100 tweets from annotator 1's batch, in order to calculate inter-annotator agreement. Using Cohen's Kappa, we found a global inter-annotator agreement of 0.504, which is seen as moderate agreement. When looking at the separate categories, we found a substantial agreement ($0.6 < \kappa < 0.8$) for *anger* ($\kappa = 0.608$) and *sadness* ($\kappa = 0.682$) and fair agreement for *fear* ($\kappa = 0.313$), *joy* ($\kappa = 0.380$) and *love* ($\kappa = 0.210$). For the *neutral* category, a moderate agreement was found ($\kappa = 0.513$).

For the dimensional annotation, we used the Best-Worst scaling approach, performed by a single trained linguist. In a previous study we conducted, we found that inter-annotator agreement was significantly higher when best-worst scaling was used compared to rating scales for labeling tweets on the dimensions *valence*, *arousal* and *dominance* (Krippendorff's alpha: 0.721, 0.349 and 0.352 for respectively *valence*, *arousal* and *dominance* in Best-Worst, versus 0.582, 0.242 and 0.112 in rating scale annotations). The 1000 tweets were converted into 2000 4-tuples, meaning that the annotator got to see 2000 trials of 4 tweets each. For each trial, the annotator had to indicate the best and worst example for each of the VAD-dimensions (i.e. highest valence and lowest valence, highest arousal and lowest arousal, and highest dominance and lowest dominance). These counts were converted to scores with the Rescorla-Wagner update rule (Rescorla et al., 1972), that assigns values to items based on the results of Best-Worst annotations. Each tweet is mapped into the three-dimensional VAD-space, using its scores obtained from the Best-Worst scaling annotation (after applying the scoring rule) as co-ordinates. A different color is used depending on the tweet's categorical annotation. For every category, the average *valence*, *arousal* and *dominance* is calculated for the tweets corresponding to that category, resulting in an average vector per category. These vectors are also drawn in the VAD-space.

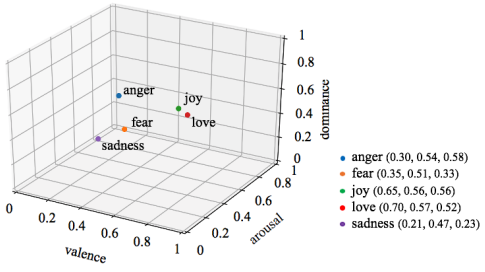
3.2.3. Results

Figure 7a shows the mapping of all instances per category mapped in the VAD-space. A clear distinction is visible between the *anger* (blue) and *joy* (green) cloud, mostly divided in terms of *valence*. In the negative *valence* area, *anger* is more or less separated from *sadness* and *fear* on the *dominance* axis. *Joy* and *love* seem to overlap rather strongly.

The separations become even clearer when looking at average vectors: for every dimension of the space, we took the average score per category and mapped these vectors in the space as well. The average VAD-scores for the five categories are shown in Figure 7b. Indeed, we see that *joy* and *love* are very close to each other in the VAD-space, while the negative emotions are better separated. *Fear* mostly dif-



(a) Mapping of tweets into the VAD-space, based on VAD-annotations. Represented in different colours depending on the tweet's categorical annotation.



(b) Mapping of emotion categories into the VAD-space, based on average *valence*, *arousal* and *dominance* of the tweet annotations.

Figure 7: Mapping of tweets into the VAD-space, based on VAD and categorical annotations.

fers from *anger* in terms of *dominance*, while *fear* and *sadness* mainly diverge on the *arousal* axis.

The image from Figure 7b is very similar to the mapping of individual emotion terms (Figure 5), with the difference that the mapping based on sentences is shifted somewhat down on the *dominance* axis and that the positive categories are placed in a more neutral area of the *valence* axis.

On the one hand, this shows that the VAD-model is rather robust (similar mapping for terms and sentences). On the other hand, the wider spread of the positive emotion categories also indicates that quite some information is lost when an emotional state is forced to be pigeon-holed in only a limited number of categories. This again strengthens the suggestion of using dimensional labels as well.

4. Conclusion

In this study, we addressed the problem of choosing an appropriate framework for building an emotion-annotated corpus. Virtually all emotion detection works fall short of giving a motivation for the framework choice, let alone an experimentally grounded one. Moreover, in the field of psychology (where these frameworks originate from), not a single study deals with fully-fledged textual data, making it unclear whether these frameworks can successfully be adopted for research in NLP.

In this work, we performed two case studies, using real-life data. Study 1 addressed categorical annotations. By means of a cluster analysis, we examined the importance of domain and topic on categorical label sets in two different domains, namely Dutch twitter messages and subtitles from reality TV shows. Study 2 explored the robustness of the VAD-model on real-life data, and examined how emotional categories relate to dimensions.

We found that the labels from the cluster analysis (namely *anger*, *fear*, *joy*, *love* and *sadness*), are well-suited to annotate text from various origins, but that the composition of the clusters and the connotation of the labels strongly depends on the domain and topic of the texts. Moreover, although this label set already has one positive emotion extra compared to Ekman, it seems that the categories *joy* and *love* cannot express all nuances of positive emotional states, indicating that information is lost when emotional states are forcedly classified in a limited set of categories. Dimensional annotations, on the other hand, seemed fairly robust, at least regarding language, rating study and text type (words versus sentences). Following Buechel and Hahn (2017), we thus believe in the advantage of a bi-representational corpus design, where categorical labels are accompanied by dimensional annotations.

These findings will help us in the creation of a Dutch Emotion Corpus, in which Dutch Tweets and subtitles from reality TV shows will be annotated both with the categorical labels *anger*, *fear*, *joy*, *love* and *sadness*, and with scores for the dimensions *valence*, *arousal* and *dominance*. Moreover, by including different domains and topics, we meet the suggestion made by Vaassen (2014) of working within a cross-domain setting.

5. Acknowledgements

This research was carried out with the support of the Research Foundation - Flanders under a Strategic Basic Research fellowship.

6. Bibliographical References

Bradley, M. M. and Lang, P. J. (1994). Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25(1):49–59.

Buechel, S. and Hahn, U. (2016). Emotion analysis as a regression problem—dimensional models and their implications on emotion representation and metrical evaluation. In *Proceedings of the Twenty-second European Conference on Artificial Intelligence*, pages 1114–1122. IOS Press.

Buechel, S. and Hahn, U. (2017). Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585.

Chatterjee, A., Narahari, K. N., Joshi, M., and Agrawal, P. (2019). SemEval-2019 task 3: EmoContext contextual emotion detection in text. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 39–48, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

De Bruyne, L., De Clercq, O., and Hoste, V. (2019). Towards an empirically grounded framework for emotion analysis. In *HUSO 2019: The Fifth International Conference on Human and Social Analytics*, pages 11–16. IARIA.

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

Fontaine, J. R., Scherer, K. R., Roesch, E. B., and Ellsworth, P. C. (2007). The world of emotions is not two-dimensional. *Psychological Science*, 18(12):1050–1057.

Leary, T. (1957). *Interpersonal diagnosis of personality: A functional theory and methodology for personality evaluation*. Ronald Press Company.

Mehrabian, A. and Russell, J. A. (1974). *An Approach to Environmental Psychology*. MIT Press.

Mohammad, S. and Bravo-Marquez, F. (2017). WASSA-2017 shared task on emotion intensity. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 34–49, Copenhagen, Denmark, September. Association for Computational Linguistics.

Mohammad, S., Bravo-Marquez, F., Salameh, M., and Kiritchenko, S. (2018). SemEval-2018 task 1: Affect in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana, June. Association for Computational Linguistics.

Moors, A., De Houwer, J., Hermans, D., Wanmaker, S., Van Schie, K., Van Harmelen, A.-L., De Schryver, M., De Winne, J., and Brysbaert, M. (2013). Norms

of valence, arousal, dominance, and age of acquisition for 4,300 dutch words. *Behavior research methods*, 45(1):169–177.

Pestian, J. P., Matykiewicz, P., Linn-Gust, M., South, B., Uzuner, O., Wiebe, J., Cohen, K. B., Hurdle, J., and Brew, C. (2012). Sentiment analysis of suicide notes: A shared task. *Biomedical informatics insights*, 5:BII–S9042.

Plutchik, R. (1980). A general psychoevolutionary theory of emotion. In Robert Plutchik et al., editors, *Theories of Emotion*, pages 3–33. Academic Press.

Rescorla, R. A., Wagner, A. R., et al. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical conditioning II: Current research and theory*, 2:64–99.

Roseman, I. J. (1984). Cognitive determinants of emotion: A structural theory. *Review of Personality & Social Psychology*, 5:11–36.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.

Schröder, M., Pirker, H., and Lamolle, M. (2006). First suggestions for an emotion annotation and representation language. In *Proceedings of LREC*, volume 6, pages 88–92.

Shaver, P., Schwartz, J., Kirson, D., and O'Connor, C. (1987). Emotion knowledge: Further exploration of a prototype approach. *Journal of personality and social psychology*, 52(6):1061–1086.

Vaassen, F. (2014). *Measuring emotion. Exploring the feasibility of automatically classifying emotional text*. Ph.D. thesis, University of Antwerp.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244.

Wood, I. D., McCrae, J. P., Andryushechkin, V., and Buitelaar, P. (2018). A comparison of emotion annotation schemes and a new annotated data set. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Languages Resources Association (ELRA).

7. Language Resource References

Alm, Cecilia Ovesdotter and Roth, Dan and Sproat, Richard. (2005). *Emotions from Text: Machine Learning for Text-based Emotion Prediction*. Association for Computational Linguistics.

Aman, Saima and Szpakowicz, Stan. (2007). *Identifying Expressions of Emotion in Text*. Springer Berlin Heidelberg.

Buechel, Sven and Hahn, Udo. (2017). *EMOBANK: Studying the impact of annotation perspective and representation format on dimensional emotion analysis*.

Figure Eight/CrowdFlower. (2016). *Emotion In Text*. Figure Eight.

Ghazi, Diman and Inkpen, Diana and Szpakowicz, Stan. (2015). *Detecting Emotion Stimuli in Emotion-Bearing Sentences*. Springer International Publishing.

Li, Yanran and Su, Hui and Shen, Xiaoyu and Li, Wenjie and Cao, Ziqiang and Niu, Shuzi. (2017). *DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset*. Asian Federation of Natural Language Processing.

Mohammad, Saif and Bravo-Marquez, Felipe. (2017). *WASSA-2017 Shared Task on Emotion Intensity*. Association for Computational Linguistics.

Saif Mohammad and Xiaodan Zhu and Svetlana Kiritchenko and Joel Martin. (2015). *Sentiment, emotion, purpose, and style in electoral tweets*.

Mohammad, Saif and Bravo-Marquez, Felipe and Salameh, Mohammad and Kiritchenko, Svetlana. (2018). *SemEval-2018 Task 1: Affect in Tweets*. Association for Computational Linguistics.

Mohammad, Saif. (2012). *#Emotional Tweets*. Association for Computational Linguistics.

Preoţiuc-Pietro, Daniel and Schwartz, H Andrew and Park, Gregory and Eichstaedt, Johannes and Kern, Margaret and Ungar, Lyle and Shulman, Elisabeth. (2016). *Modelling valence and arousal in facebook posts*.

Scherer, Klaus R and Wallbott, Harald G. (1994). *Evidence for universality and cultural variation of differential emotion response patterning.* American Psychological Association.

Schuff, Hendrik and Barnes, Jeremy and Mohme, Julian and Padó, Sebastian and Klinger, Roman. (2017). *Annotation, Modelling and Analysis of Fine-Grained Emotions on a Stance and Sentiment Detection Corpus*. Association for Computational Linguistics.

Strapparava, Carlo and Mihalcea, Rada. (2007). *SemEval-2007 Task 14: Affective Text*. Association for Computational Linguistics.

Vaassen, Frederik and Daelemans, Walter. (2011). *Automatic Emotion Classification for Interpersonal Communication*. Association for Computational Linguistics.