Computational Linguistics in the Netherlands Journal 11 (2021) 231-255 Submitted 10/2021; Published 12/2021

Prospects for Dutch Emotion Detection: Insights from the New EmotioNL Dataset

Luna De Bruyne LUNA.DEBRUYNE@UGENT.BE
Orphée De Clercq ORPHEE.DECLERCQ@UGENT.BE
Véronique Hoste VERONIQUE.HOSTE@UGENT.BE

LT³, Language and Translation Technology Team, Ghent University

Abstract

Although emotion detection has become a crucial research direction in NLP, the main focus is on English resources and data. The main obstacles for more specialized emotion detection are the lack of annotated data in smaller languages and the limited emotion taxonomy. In a first step towards improving emotion detection for Dutch, we present EmotioNL, an emotion dataset consisting of 1,000 Dutch tweets and 1,000 captions from TV-shows, annotated with emotion categories (*anger*, *fear*, *joy*, *love*, *sadness* and *neutral*) and dimensions (*valence*, *arousal* and *dominance*). We evaluate the state-of-the-art Dutch transformer models BERTje and RobBERT on this new dataset, investigate model generalizability across domains and perform a thorough error analysis based on the Component Process Model of emotions.

1. Introduction

Emotion detection has become a crucial research direction in Natural Language Processing (NLP). Although recent studies have shown interest in multilingual emotion detection (Buechel and Hahn 2018, Öhman et al. 2018) or emotion detection for different languages (Ahmad et al. 2020), the English language still receives most attention. Algorithmic breakthroughs like the use of bi-directional LSTMs, attention and the state-of-the-art transformer models (Devlin et al. 2019) have fueled optimism in artificial emotional intelligence being within reach, exemplified by promising results in the past SemEval competitions (Chatterjee et al. 2019, Mohammad et al. 2018). However, these well-performing systems primarily concern the detection of basic emotions like *anger*, *fear*, *sadness* and *joy* and are mostly restricted to English. For smaller languages such as Dutch, a lot of ground still has to be covered.

As brought up by Vaassen (2014), who was the first to focus on Dutch emotion detection, the main obstacle for emotion classification is the subjective nature of the task, and the ensuing lack of high-quality annotated datasets. He also questions the feasibility of constructing large, high-agreement emotion datasets, and therefore advocates to collect data from different domains which could help to increase agreement (e.g. subtitles, as they also incorporate a visual component). Furthermore, instead of focusing too much on correct versus incorrect predictions, he suggests to rethink the notion of a gold standard and introduces the idea of an acceptability scale. Similarly, Buechel and Hahn (2016) called to treat emotion analysis as a regression problem rather than a classification task.

Given that the only publicly available Dutch emotion-annotated dataset, deLearyous (Vaassen and Daelemans 2011), only comprises 740 instances and exhibits a low inter-annotator agreement, there is a clear need for Dutch corpora annotated with emotions, preferably containing multiple domains and emotion representations. To this purpose we present EmotioNL, an emotion dataset that consists of 2,000 textual instances in two domains: 1,000 Dutch Twitter messages and 1,000 captions from Flemish reality TV-shows. The data has been annotated in a bi-representational format: both with emotional categories (*anger*, *fear*, *joy*, *love*, *sadness* and *neutral*) and emotional dimensions (scores for *valence*, *arousal* and *dominance*). Annotator agreement was ensured by performing a cluster analysis to establish the categorical label set for the first format (De Bruyne et al. 2020) and by investigating different annotation methods for assigning the emotional dimension scores (De Bruyne et al. 2021a).

In line with Vaassen's above-mentioned idea of an acceptability scale, we also introduce a new metric called 'cost-corrected accuracy'. Although the gold standard is still accepted, this metric takes into account the severity of a false prediction (or the 'cost'). For example, misclassifying an instance of *joy* as *love* is a less severe mistake than misclassifying that same instance as *anger*. This will account for some of the variability between annotators and, more than that, allows for a fairer evaluation of emotion classification models.

Although the size of EmotioNL might still be considered rather small, it is already a big improvement compared to Delearyous (i.e. 2,000 versus 740 instances). Moreover, the proposed datasets are suitable for testing transfer learning techniques. In this respect, transformer models are currently considered state-of-the-art for many NLP tasks (Devlin et al. 2019), and brought forth Dutch models like BERTje (de Vries et al. 2019) and RobBERT (Delobelle et al. 2020). Until now, however, the (limited) research on Dutch emotion detection is still focused on linear classifiers and lexicon-based approaches (Vaassen 2014) whereas this dataset will enable fine-tuning transformer models on emotional data as well.

Furthermore, EmotioNL lends itself well for investigating cross-domain generalizability. Our previous research (De Bruyne et al. 2020) already suggested that emotional connotation can change according to the domain. This is why in this work various experiments are performed to investigate whether training on both domains has an added value compared to training solely on the (smaller) target domain and whether a model trained on one domain can be used for inference on the other domain.

The main contributions of this paper are a) presenting a new high-quality dataset with Dutch emotion annotations for two domains in a bi-representational format[1], b) introducing a new metric, named cost-corrected accuracy, which takes into account the misclassification cost, c) investigating cross-domain transfer learning and d) providing a thorough error analysis to get more insights into the challenges and possibilities of Dutch emotion detection.

We will discuss related work on Dutch emotion detection in Section 2, as well as research on cross-domain emotion detection and emotion detection metrics. In Section 3, the design and creation of the dataset are described. Section 4 deals with the evaluation of the dataset and describes the experimental set-up of some baseline emotion detection and cross-domain transfer experiments (Section 4.1), introduces the concept of cost-corrected accuracy (Section 4.2) and reports the experimental results (Section 4.3). Section 5 will be dedicated to an error analysis, and we end this paper with a conclusion in Section 6.

2. Related work

2.1 Dutch emotion detection

Most studies in the field of sentiment and emotion analysis dealing with the Dutch language are restricted to polarity analysis (negative-neutral-positive) and do not involve the identification of fine-grained emotions. The first efforts in Dutch sentiment analysis primarily focused on the creation of affect lexicons to be leveraged in keyword-based approaches. One of the first Dutch sentiment lexica is DuOMAn (Jijkoun and Hofmann 2009), which was generated based on translations of an English lexicon and lexical relations like synonymy and antonymy. Another popular resource is the subjectivity lexicon included in the Pattern[2] library (De Smedt and Daelemans 2012). Both lexicons are being used for conducting sentiment analysis in Dutch to this very day, e.g. DuOMAn was used for sentiment analysis in financial news (Van de Kauter et al. 2015) and Pattern for research analyzing the sentiment in Tweets about governmental measures against COVID-19 (Wang et al. 2020).

In a move towards more fine-grained sentiment analysis, aspect-based sentiment analysis started to hit the spotlight. This task not only deals with identifying sentiment expressions, but also the concepts (or aspects) to which they refer. De Clercq et al. (2017) were among the first to develop an integrated pipeline to solve this task in various domains for Dutch.

Perhaps even more complex is the shift from sentiment to emotion. Already in the early 2000s, the first research on emotion detection started for English (Holzman and Pottenger 2003, Alm et al. 2005). This

[1] Dataset available on request by contacting luna.debruyne@ugent.be.

[2] http://www.clips.ua.ac.be/pages/pattern

attracted attention in the Dutch language area as well, although the number of studies on this topic remained rather limited. The first researchers studying emotion detection in Dutch were Vaassen and Daelemans (2011), who attempted to classify conversational sentences according to Leary's Rose or the Interpersonal Circumplex (Leary 1957). To this purpose a dataset comprising 740 Dutch sentences with annotations for the octants of Leary's Rose was established. SVM-based classifiers achieved macro-averaged F1-scores of up to 31% for 8-way classification into the octants, and 51% F1-score for 4-way classification into the quadrants of the Rose. Later, the authors developed a dataset containing 11 conversations annotated on the sentence level (1,143 sentences) with their position on Leary's Rose and made this dataset publicly available as the deLearyous dataset (Vaassen et al. 2012).

Tromp and Pechenizkiy (2014) employed a rule-based method for detecting the eight basic emotions of Plutchik (1980) (*anger, fear, joy, sadness, anticipation, surprise, disgust* and *trust*) in multiple datasets, including a Dutch Twitter data set consisting of 402 Dutch tweets. Their pattern and ruled-based algorithm reached an accuracy of 56.7%.

In the context of emotion analysis in Dutch historical texts, the Historic Embodied Emotion Model (HEEM) was developed by van der Zwaan et al. (2015). They constructed a dataset of 29 Dutch 17th and 18th century theatre plays, manually annotated with 38 emotions labels and additional annotations for body parts, bodily processes, emotional actions and body sensations involved in the emotion expressed in the text, with the aim of tracing historical changes in the verbal and bodily expression of emotions over time. Binary Relevance and Random k-Labelsets were used with Linear Support Vector Machines as classification algorithms, which reached a micro and macro-averaged F1-score of 0.45 and 0.24 respectively.

Finally, some researchers have attempted to automatically analyse emotions using the Dutch version of the Linguistic Inquiry and Word Count or LIWC (Boot et al. 2017), e.g. to analyse emotional differences between texts containing socially acceptable and unacceptable discourse (Ljubešić et al. 2020), to track emotional expressions on Twitter after the outbreak of the COVID-19 pandemic (Metzler et al. 2021), to analyse emotions in parliamentary war victim debates (van Lange and Futselaar 2021) or to examine gender differences in the emotional language in the Dutch Veteran Institute Oral History Archive (Roumen 2021). However, these studies are not concerned with studying automatic emotion detection itself, but merely apply existing methods.

In the meantime, research on English emotion detection kept on developing. Especially with the advent of Transformer models like BERT (Devlin et al. 2019), emotion detection research flourishes (Graterol et al. 2021). For Dutch, transformer models have also been developed, like the BERT-based BERTje (de Vries et al. 2019) and RoBERTa-based RobBERT (Delobelle et al. 2020). Although they achieved promising results on the task of sentiment analysis, these models have not yet been evaluated on emotional data. The only study in which Dutch transformer models have been used for Dutch emotion detection, is our own work described in De Bruyne et al. (2021b). These were preliminary experiments which investigated the possibility to combine transformer architectures with affect lexica and were only tested on a small subset of the data described in the current paper.

2.2 Cross-domain emotion detection

A well-known problem in sentiment and emotion analysis, is polarity divergence (Zhang et al. 2015). This means that words, depending on a particular context or domain, might have a different polarity. In previous work, we also found that the interpretation and connotation of emotion-related concepts is domain-dependent, and this not only at the lexical level, but even at the level of the emotion labels themselves (De Bruyne et al. 2020). Moreover, like in any NLP task, different domains might have different distributions of labels, resulting in models which are difficult to generalize across domains.

The subject of domain adaption, which deals with developing techniques to make models generalizable over domains, is widely studied in NLP. Originally, domain adaptation techniques were focused on modelling the relationship between the distribution of the in-domain and out-of-domain data (Daume III and Marcu 2006). A simple domain adaption technique was proposed by Daumé III (2007), which involves feature augmentation such that the data contain general, source-specific and target-specific versions of the features.

The advent of transformer models led to a new take on domain adaptation and transfer learning techniques, as transfer learning is an inherent part in the way transformer models are currently used, namely by pre-training large models and fine-tuning them on the task at hand. We could thus say that the domain adaptation occurs in the fine-tuning part. However, one could even go one step further and implement domain adaptation mechanisms in the fine-tuning architecture. One example is domain adaptation through adversarial training, which involves training a secondary task to discriminate between pre-defined domains. That way, domain-independent features can be identified, making the model robust across the different domains in the data (Du et al. 2020).

Vaassen (2014) also looked into the generalizability of his emotion detection models. He found that building a domain-independent emotion detection model is extremely difficult, but, that a model which is trained on a mixture of in-domain and out-of-domain data (in which the out-of-domain data is better balanced) can increase performance on the target data, when the in-domain data is sparse and imbalanced.

2.3 Emotion detection metrics

Given the highly subjective nature of the task of emotion annotation, Vaassen (2014) suggested to rethink the concept of a gold standard annotation. As it is almost impossible to assign an objectively correct emotion class to a piece of text, he proposed the idea to judge a classifier's output on an acceptability scale, rather than judging it as either correct or incorrect.

Although we will still maintain a gold standard as point of reference, we do believe there are some issues with current evaluation metrics used for emotion classification, like F1-score or accuracy. After all, not all mistakes are equally wrong. Classifying an instance that was annotated with *fear* as *sadness*, is a less severe mistake than classifying that instance as *joy*. To some extent, these 'misclassifications' follow the same tendencies as the variability between annotators and are often situated within the same polarity (e.g. confusing *sadness* and *fear* or *love* and *joy*). This leads to the need of a metric which also takes polarity into account, in which case we are dealing with a kind of ordinal classification problem.

One could therefore turn to regression metrics like Mean Squared Error or Mean Average Error. However, these metrics assume that all classes are equidistant, which is not necessarily the case when dealing with sentiment or emotion analysis. Also correlation metrics like Pearson's or Spearman's r are not most appropriate, as classification predictions can have a perfect correlation without any of the predictions actually matching the gold label.

Other metrics for ordinal classification were among others proposed by Waegeman et al. (2006), who extend ROC curve analysis principles to what they call ordinal regression, Amigó et al. (2020), whose *closeness evaluation measure* is based on the idea that two items a and b are informationally close if the probability of finding an item between the two is low, and George et al. (2016), who propose a metric that considers the optimal trade-off between accuracy (Acc) and misclassification cost (MC), which they define as the distance between a classifier's (Acc, MC) coordinate and the ideal coordinate of $(1, 0)$.

3. EmotioNL Design

3.1 Data collection

We collected data for a new Dutch emotion corpus, EmotioNL, from two domains, namely Twitter posts and reality TV-show captions. Both datasets consist of 1,000 instances, amounting to a corpus comprising 2,000 items.

The tweets were scraped using the Dutch database Twiqs.nl[3]. In order to increase the chance of collecting emotionally charged posts, emojis were used as search query (72 face type emojis). As such, a one-year datadump was obtained (search period from 1-1-2017 to 31-12-2017), from which a random subset of 1,000 tweets was sampled, excluding duplicates or tweets in other languages than Dutch. The emojis are preserved in the dataset, but can of course be removed if desired at a later stage.

[3] http://145.100.59.103/cgi-bin/twitter

	V	A	D
rating scale	0.564	0.208	0.119
pairwise comparison	0.580	0.265	0.143
best-worst scaling	0.709	0.388	0.360

Table 1: IAA (Krippendorff's α) of different rating methods (rating scale, pairwise comparison and best-worst scaling) on the dimensions valence (V), arousal (A) and dominance (D). Results from De Bruyne et al. (2021a).

For the Captions subcorpus three emotionally loaded Flemish reality TV-shows were selected, namely *Blind getrouwd*; *Bloed, zweet en luxeproblemen* and *Ooit vrij*. In *Blind getrouwd*, singles get matched by experts and get married before having seen each other; *Bloed, zweet en luxeproblemen* shows the travel story of six adolescents and their confrontation with inequality in Asia and Africa; and in *Ooit vrij*, we follow detainees on their way to release. Three episodes per show were transcribed using a literal transcription method (without correcting colloquial elements), which means that also non-standard language data is included. From these transcripts 1,000 utterances (sentences or short sequences of sentences) were selected, based on a rough screening of emotional content and an approximately equal distribution over the three shows (335 instances from *Blind getrouwd*, 331 from *Bloed, zweet en luxeproblemen* and 334 from *Ooit vrij*).

3.2 Annotation procedure

Thorough preliminary research motivated us to annotate the data in a bi-representational format, using both categorical labels and dimensional annotations, as also proposed by Buechel and Hahn (2016). Own exploratory research suggested that emotional nuances are lost when affective states are forcedly classified in a limited set of categories, reinforcing the idea of accompanying categorical labels by dimensional annotations (De Bruyne et al. 2020).

For the categorical labelling, each instance was labelled with one out of six labels: *anger*, *fear*, *joy*, *love*, *sadness* or *neutral*. This label set was based on clustering experiments, in which we aimed to obtain an experimentally grounded label set (De Bruyne et al. 2020). Starting from 25 categories, which we used to annotate 300 sentences from the Tweets corpus and 300 sentences from the Captions corpus using a multi-label approach, a cluster analysis was performed. For both the Tweets and Captions subcorpora, this cluster analysis led to a label set containing five groups clustered around *joy*, *love*, *anger*, *fear* and *sadness*. However, the composition of the clusters depended on the domain.

The dimensional annotation was based on the VAD-model by Mehrabian and Russell (1974). According to this model, each emotional state can be represented by the emotional dimensions *valence*, *arousal* and *dominance*. The usage of rating scales to annotate dimensional properties has several drawbacks, including divergent interpretations of the scale and scale bias (Kiritchenko and Mohammad 2017). We therefore used best-worst scaling as rating method, which was found to be the most reliable rating approach compared to rating scales and pairwise comparison (De Bruyne et al. 2021a). The results of that study, where Krippendorff's α was calculated to assess the inter-annotator agreement (IAA) between 6 annotators for 300 Tweets, is shown in Table 1.

Two annotators executed the labelling task. For the categorical labelling, each annotator labelled 500 distinct tweets and 500 captions. Additionally, one annotator labelled 100 items per subcorpus overlapping with the other annotator in order to calculate inter-annotator agreement. For the dimensional annotation, the 1,000 instances in each subcorpus were converted into 2,000 4-tuples and distributed among the annotators. For each trial, the annotator had to indicate the best and worst example for each dimension: highest and lowest *valence*, highest and lowest *arousal*, and highest and lowest *dominance*. Best-worst counts were then converted to scores from 0 to 1 with the Rescorla-Wagner update rule (Rescorla et al. 1972). The full annotation guidelines for the categorical and dimensional annotation are included in the Appendix.

In the Tweets subset, emojis are kept in the dataset both for annotation and further analyses. In the Captions set, some utterances were embedded in their context (e.g. in conversations), but this context was left out for further analysis. Table 2 shows an annotated example of one instance per domain.

Corpus	Text example	categorical	dimensional		
			V	A	D
Tweets	@transavia Jaaah 🙂 volgende vakantie Barcelona en na het zomerseizoen naar de Algarve *EN: @transavia Yeah 🙂 next holiday Barcelona and after summer season to the Algarve*	joy	0.689	0.491	0.622
Captions	Ik zou liever sterven dan hier te wonen, denk ik. *EN: I'd rather die than live here, I think.*	sadness	0.156	0.384	0.301

Table 2: Text examples from the Tweets and Captions subcorpora with their assigned categorical and dimensional label (V = valence, A = arousal, D = dominance).

3.3 Corpus statistics

We found a global IAA (Cohen's Kappa) of 0.504 for emotion categories in Tweets and 0.568 in Captions, which is seen as moderate agreement. Agreement for separate categories is shown in Table 3. In the Tweets subset, 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 Captions, substantial agreement was found for *anger* ($\kappa = 0.740$) and *joy* ($\kappa = 0.721$) and fair agreement for *fear* ($\kappa = 0.497$), *love* ($\kappa = 0.370$), *sadness* ($\kappa = 0.428$) and *neutral* ($\kappa = 0.470$). For all categories except *sadness* and *neutral*, IAA is higher in Captions than in Tweets, possibly because the annotators disposed of more context about the captions (annotations and transcriptions were done by the same people, so annotators had watched the video material and thus also had visual context).

Figures 1 and 2 visualise the confusion between the annotations of both annotators for the Tweets and Captions subsets respectively. In Tweets, most disagreement is between the pairs *joy–neutral* and *joy–love*, while in Captions, the most notable disagreement was between *sadness* and *neutral*.

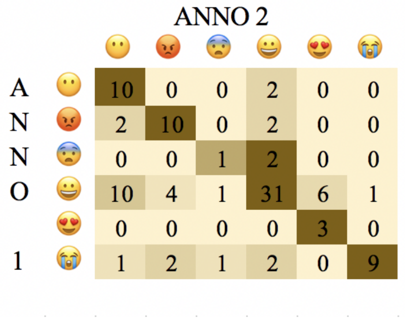
	A	F	J	L	S	N
Tweets	.608	.313	.380	.210	.682	.513
Captions	.740	.497	.721	.370	.428	.470

Table 3: IAA (κ) for each emotion category per subdataset.

	Tweets	Captions
Anger	188	198
Fear	51	96
Joy	400	340
Love	44	45
Sadness	98	186
Neutral	219	135

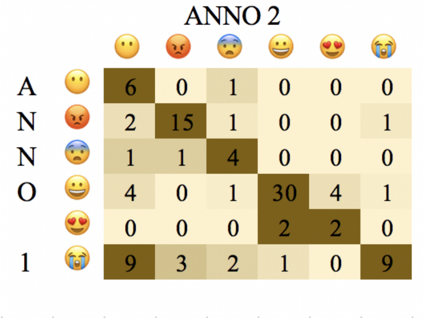
Table 4: Number of instances in each emotion category per subdataset.

Table 4 shows the number of instances per emotion category in each domain. For the *valence*, *arousal* and *dominance* annotations, mean ranges between 0.46 and 0.52 for all dimensions in both subsets, standard



	🙂	😡	😨	😀	😍	😭
🙂	10	0	0	2	0	0
😡	2	10	0	2	0	0
😨	0	0	1	2	0	0
😀	10	4	1	31	6	1
😍	0	0	0	0	3	0
😭	1	2	1	2	0	9

Figure 1: Confusion matrix between annotators for Tweets.
🙂 = neutral; 😡 = anger; 😨 = fear; 😀 = joy; 😍 = love; 😭 = sadness.



	🙂	😡	😨	😀	😍	😭
🙂	6	0	1	0	0	0
😡	2	15	1	0	0	1
😨	1	1	4	0	0	0
😀	4	0	1	30	4	1
😍	0	0	0	2	2	0
😭	9	3	2	1	0	9

Figure 2: Confusion matrix between annotators for Captions.
🙂 = neutral; 😡 = anger; 😨 = fear; 😀 = joy; 😍 = love; 😭 = sadness.

	Tweets			Captions		
	V	A	D	V	A	D
Mean	0.50	0.51	0.50	0.46	0.48	0.52
SD	0.22	0.20	0.20	0.19	0.20	0.18
Min	0.06	0.06	0.06	0.07	0.05	0.05
Max	0.97	0.97	0.96	0.96	0.96	0.97

Table 5: Mean, standard deviation, minimum and maximum for the distributions of valence, arousal and dominance per subdataset.

deviation between 0.18 and 0.22, minimum between 0.05 and 0.07 and maximum between 0.96 and 0.97 (see Table 5).

4. Evaluation

4.1 Cross-domain transfer: Experimental set-up

Transformer models are considered state-of-the-art for many NLP tasks (Devlin et al. 2019, Liu et al. 2019). For Dutch, both a BERT-based model, namely BERTje (de Vries et al. 2019) and a RoBERTa-based model, RobBERT (Delobelle et al. 2020), have been developed. The latter achieved state-of-the-art performance on a sentiment analysis benchmark dataset of Dutch book reviews (Van der Burgh and Verberne 2019). With EmotioNL, these models can be evaluated on the task of emotion analysis as well.

The main differences between BERTje and RobBERT is that the former is a BERT-based model, while the latter uses the RoBERTa methodology and is trained on much more data than BERTje. While BERTje uses the BERT pre-training tasks of Masked Language Model (MLM) and Next Structure Prediction (NSP) to learn a language representation, RobBERT follows the RoBERTa methodology by dropping the NSP task and using dynamic masking instead. As regards the pre-training data, BERTje uses 12GB of data coming from novels, Wikipedia, news from the TwNC corpus (Ordelman et al. 2007) and web news, and the multi-genre reference corpus SoNaR-500 (Oostdijk et al. 2013). RobBERT, on the other hand, uses 39GB of data, coming from the Dutch section of the Common Crawl corpus (Suárez et al. 2019).

First, model selection experiments are performed in order to compare the performance of BERTje and RobBERT on both domains and tasks (classification and regression) in the EmotioNL corpus. Parameter settings are shown in Table 6. The hyperparameters are based on the parameters reported in the papers of

Optimizer	AdamW
Learning rate	$5e-5$
Learning rate scheduler	ReduceLROnPlateau
Loss function	BCE (classification) and MSE (regression)
Activation function	GELU
Dropout	0.2
Max sequence length	64
Batch size	64
Epochs	5 (classification) and 10 (regression)
GPU	Nvidia Tesla V100

Table 6: Parameters and settings for BERTje and RobBERT.

Devlin et al. (2019), Lio et al. (2019), de Vries et al. (2019) and Delobelle et al. (2020) (e.g. learning rate and dropout), computational capacity of our GPUs (sequence length and batch size) and some preliminary experiments (number of epochs).

As EmotioNL contains data for two domains, we investigate its cross-domain generalizability. We therefore foresee three different setups (cf. Figure 3): in the first one (setup 1), we train on each domain separately and test on in-domain data (cf. model selection experiments); next, we train on one domain, and test on the out-of-domain data (setup 2); finally, we train on both datasets and test the performance on each domain separately (setup 3).

All models are evaluated using 10-fold cross validation. For the model selection and setup 1 experiments, this means that there are 10 models that are each time trained on 900 instances and tested on 100 instances from the same domain. In setup 3, we train on all instances from the source domain and 900 from the target domain, and test on the remaining 100 from the target domain, and repeat this for each fold (once where Captions is the target domain and once where Tweets is the target domain). In setup 2, cross-validation is not needed, as we train on the full source dataset and test on the full target dataset in one go.

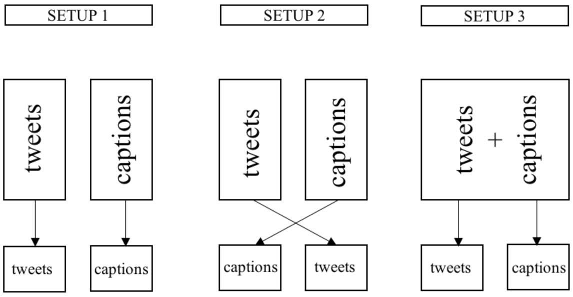


Figure 3: Different setups for the cross-domain experiments. Setup 1: training and testing on in-domain data; Setup 2: testing on out-of-domain data; Setup 3: training on in and out-of domain data.

4.2 Metrics

4.2.1 Classification metrics

Commonly used metrics for evaluating multi-class classification tasks are accuracy and macro F1. Macro-averaged F1 score is the harmonic mean of precision and recall, in which each class is treated equally and thus compensates for class imbalance. In multi-class classification, accuracy is equal to micro F1 (which does not compensate for class imbalance).

Although both accuracy and F1 are widely used metrics, we think that, specifically for emotion classification, these metrics fall short. We believe that not all mistakes are equally wrong, and that the severity of misclassifications needs to be taken into account when evaluating models.

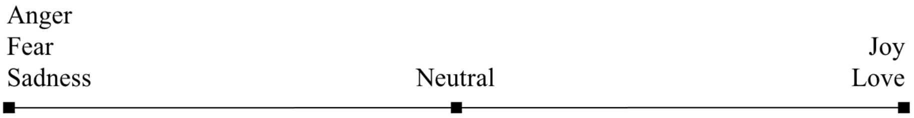


Figure 4: Ordinal placement of the emotion categories *anger*, *fear*, *joy*, *love*, *sadness* and *neutral*.

	anger	fear	sadness	neutral	joy	love
anger	0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{2}{3}$	1	1
fear	$\frac{1}{3}$	0	$\frac{1}{3}$	$\frac{2}{3}$	1	1
sadness	$\frac{1}{3}$	$\frac{1}{3}$	0	$\frac{2}{3}$	1	1
neutral	$\frac{2}{3}$	$\frac{2}{3}$	$\frac{2}{3}$	0	$\frac{2}{3}$	$\frac{2}{3}$
joy	1	1	1	$\frac{2}{3}$	0	$\frac{1}{3}$
love	1	1	1	$\frac{2}{3}$	$\frac{1}{3}$	0

Figure 5: Cost matrix with adjusted weights for misclassifications, based on polarity.

In line with George et al. (2016), we propose a metric based on accuracy and misclassification cost and call it ‘cost-corrected accuracy’. We interpret the misclassification cost in terms of polarity, where misclassification within the same polarity (e.g. classifying an instance of *love* as *joy*) is punished less than a misclassification with an opposite polarity (e.g. classifying an instance of *love* as *anger*).

Therefore, we place our emotion labels on an ordinal scale, with *anger*, *fear* and *sadness* on the lower end, *neutral* in the middle, and *joy* and *love* on the upper end (see Figure 4). Based on this scale, a cost matrix is defined where a correct prediction has a cost of 0, a misclassification with an opposite polarity a cost of 1 and other misclassifications a cost of 1/3 or 2/3, depending on the divergence (Figure 5).

The total cost can then be calculated by multiplying the confusion matrix (Cf) with the cost matrix (Ct):

$$\mathit{cost} = Cf \bullet Ct \tag{1}$$

The lower the cost, the better the model’s performance. Because our weights have a maximum value of 1, the maximum cost will be 1 and the minimal cost will be 0. However, to turn this into an accuracy-like measure, we then define cost-corrected accuracy ($CC\text{-}ACC$) as:

$$CC\text{-}ACC = 1 - \mathit{cost} \tag{2}$$

Note that in normal accuracy, the same happens. The only difference there is that all values in the cost matrix are either 0 or 1. However, by adjusting the cost weights, the ordinal nature of emotion classification is taken into account.

Another possibility, which we will not employ in this paper, is to put extra weight on certain classes (e.g. to account for class imbalance) by increasing their cost value in the cost matrix. As cost-corrected accuracy is based on accuracy, it is sensitive to class imbalance.

4.2.2 Regression metrics

For evaluating the prediction of the VAD values, which is a regression task, we will use Pearson's correlation coefficient r. We calculate the correlation between gold standard and predicted value for each dimension, and then take the average.

4.3 Results

4.3.1 Model selection experiments

First, we evaluate BERTje and RobBERT on both domains. Tables 7 and 8 show the results for the classification and regression task, respectively. Except for Captions classification, RobBERT clearly outperforms BERTje. Especially for Tweets regression, BERTje performs notably worse than RobBERT. However, for Captions classification, BERTje outperforms RobBERT, be it only to a small extent. Therefore, we will use RobBERT for all further experiments.

As can be observed, the overall results for these tasks are low. This is especially the case for the classification tasks, where macro-F1 and accuracy are only around 35% and 50% respectively. Cost-corrected accuracy, which takes into account the severity of misclassifications, reveals more optimistic scores of 69% for Tweets and 65% for Captions. Performance on the regression tasks (i.e. emotional dimensions), which exhibit a Pearson's correlation coefficient of around 64%, seems more promising.

	Tweets			**Captions**		
Model	**F1**	**Acc.**	**Cc-Acc.**	**F1**	**Acc.**	**Cc-Acc.**
BERTje	0.257	0.430	0.573	0.381	0.496	0.662
RobBERT	0.347	0.539	0.692	0.372	0.478	0.654

Table 7: Macro F1, accuracy and cost-corrected accuracy for the different models on the classification task in the Tweets Captions subset.

	Tweets	**Captions**
Model	r	r
BERTje	0.238	0.548
RobBERT	0.635	0.641

Table 8: Pearson's r for the different models on the VAD regression task in the Tweets and Captions subset.

4.3.2 Cross-domain experiments

In these sets of experiments, the generalizability of our models across the domains in EmotioNL is scrutinized. Setup 1 is equal to the experiments with RobBERT in the previous section, as we train and test on in-domain data. In setup 2, we train a model on a source domain and test it on the other domain (i.e. target domain). In setup 3, we train on both the source (full) and target (10 times 900 instances) domain and test it on the target domain. The results are presented in Tables 9 and 10, where the left-hand side each time represents the scenario with Tweets as target domain and the right-hand side Captions as target.

Intuitively, one would expect the performance to decrease in the cross-domain setup (setup 2) compared to the in-domain setting (setup 1), as there is probably a discrepancy between source and target domain. Indeed, for all tasks and metrics a clear drop in performance can be perceived in both domains: for classification, (cost-corrected) accuracy decreases with around 10%, and for regression a drop of more than 20% of the Pearson correlation coefficient is observed.

Model	Tweets F1	Tweets Acc.	Tweets Cc-Acc.	Captions F1	Captions Acc.	Captions Cc-Acc.
Setup 1: in-domain	0.347	0.539	0.692	0.372	0.478	0.654
Setup 2: cross-domain	0.283	0.430	0.611	0.274	0.352	0.574
Setup 3: multi-domain	0.381	0.515	0.670	0.396	0.517	0.677

Table 9: Macro F1, accuracy and cost-corrected accuracy for the different models on the classification task in the Tweets and Captions subset.

Model	Tweets r	Captions r
Setup 1: in-domain	0.635	0.641
Setup 2: cross-domain	0.422	0.437
Setup 3: multi-domain	0.653	0.624

Table 10: Pearson's r for the different models on the VAD regression task in the Tweets and Captions subset.

In setup 3, where we train on both domains, performance goes up for Captions classification (an increase of around 2%) and Tweets regression, but it decreases in Captions regression (with 2%) . In Tweets classification, the multi-domain setting causes an increase in terms of macro F1 (from 35% to 38%), but a small decrease in (cost-corrected) accuracy (drop of 2%). Mc-Nemar's tests for evaluating the difference in performance between the in-domain and multi-domain setting indicate that the multi-domain is significantly better than the in-domain setting in Captions (p = 0.016), but not for Tweets (p = 0.158). There is thus no undisputed positive effect of training on two domains (see Section 5.2 for a discussion in closer detail).

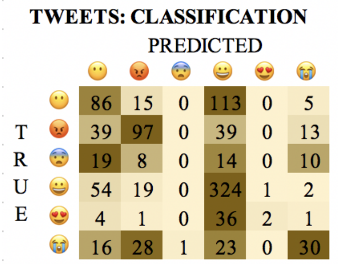
5. Discussion

In this section, we wish to give more insights into the predictions made by the models. We will first look at the confusion matrices of the in-domain RobBERT models and then zoom in on the different emotion representations and tasks (classification versus regression, Section 5.1) and the two domains in our corpus (Tweets versus Captions, Section 5.2). Finally, we will dive into the textual instances themselves and look for possible challenges for emotion prediction (Section 5.3). For this latter analysis, we will turn to the Emotion Component Process model by Scherer (2005).

5.1 Classification versus regression

Confusion matrices for the classification tasks (in-domain RobBERT) are shown in Figures 6 and 7. These clearly show the dire performance on the minority classes *love* and (especially in Tweets) *fear*. In the Tweets domain, many instances are wrongly classified as *joy* or (but less often) as *neutral*, and the negative classes are mixed up. In the Captions domain, more or less the same trend can be observed, though less pronounced. This is mainly reflected in the higher macro F1 for Captions (.37 versus .35), even though accuracy is lower in this domain than in Tweets (.47 versus .54).

For regression, confusion is visualised by grouping the VAD scores in 4 categories: below .2 (= very low *valence* / *arousal* / *dominance*), between .2 and .4 (= low *valence* / *arousal* / *dominance*), between .4 and .6 (= moderate *valence* / *arousal* / *dominance*), between .6 and .8 (= high *valence* / *arousal* / *dominance*) and above .8 (= very high *valence* / *arousal* / *dominance*). The corresponding confusion matrices are shown in Figures 8 and 9.

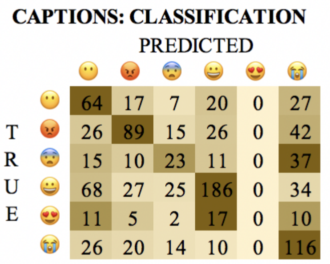


TWEETS: CLASSIFICATION

PREDICTED

		🙂	😡	😨	😀	😍	😭
	🙂	86	15	0	113	0	5
T	😡	39	97	0	39	0	13
R	😨	19	8	0	14	0	10
U	😀	54	19	0	324	1	2
E	😍	4	1	0	36	2	1
	😭	16	28	1	23	0	30

Figure 6: Confusion matrix Tweets classification. 🙂 = neutral; 😡 = anger; 😨 = fear; 😀 = joy; 😍 = love; 😭 = sadness.



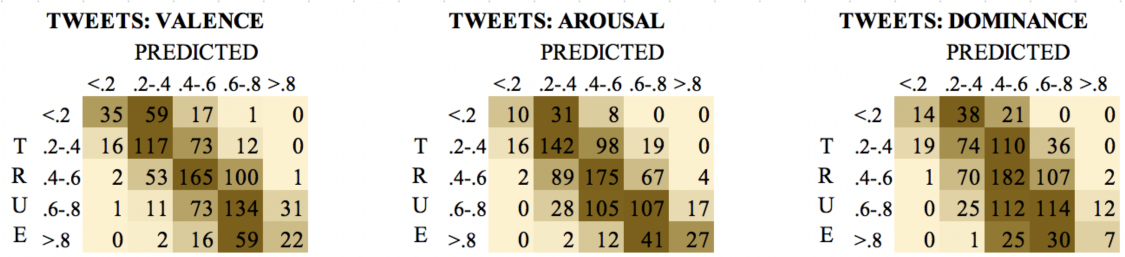
CAPTIONS: CLASSIFICATION

PREDICTED

		🙂	😡	😨	😀	😍	😭
	🙂	64	17	7	20	0	27
T	😡	26	89	15	26	0	42
R	😨	15	10	23	11	0	37
U	😀	68	27	25	186	0	34
E	😍	11	5	2	17	0	10
	😭	26	20	14	10	0	116

Figure 7: Confusion matrix Captions classification. 🙂 = neutral; 😡 = anger; 😨 = fear; 😀 = joy; 😍 = love; 😭 = sadness.

Although the predictions are to a great extent clustered around the diagonal (which corresponds to a high degree of true positives), we see that the model is hesitant to predict the most extreme classes (below .2 or above .8), especially for *dominance*. However, all in all, the confusion is less marked for the regression than for the classification task. It does seem that predicting emotional categories is more challenging than predicting scores for emotional dimensions, which supports the claims made by Buechel and Hahn (2016).



TWEETS: VALENCE

PREDICTED

		<.2	.2-.4	.4-.6	.6-.8	>.8
	<.2	35	59	17	1	0
T	.2-.4	16	117	73	12	0
R	.4-.6	2	53	165	100	1
U	.6-.8	1	11	73	134	31
E	>.8	0	2	16	59	22

TWEETS: AROUSAL

PREDICTED

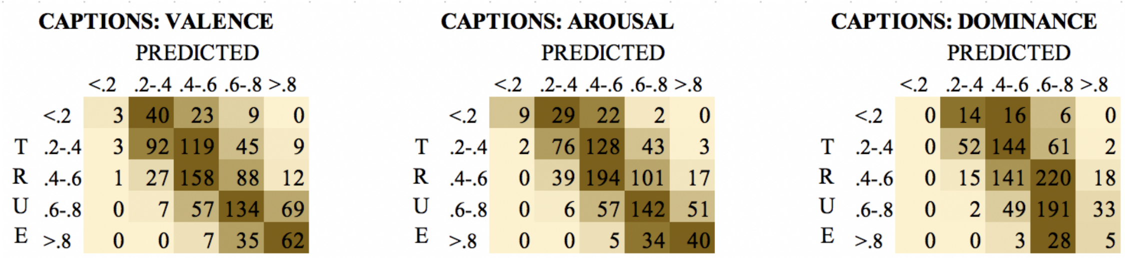
		<.2	.2-.4	.4-.6	.6-.8	>.8
	<.2	10	31	8	0	0
T	.2-.4	16	142	98	19	0
R	.4-.6	2	89	175	67	4
U	.6-.8	0	28	105	107	17
E	>.8	0	2	12	41	27

TWEETS: DOMINANCE

PREDICTED

		<.2	.2-.4	.4-.6	.6-.8	>.8
	<.2	14	38	21	0	0
T	.2-.4	19	74	110	36	0
R	.4-.6	1	70	182	107	2
U	.6-.8	0	25	112	114	12
E	>.8	0	1	25	30	7

Figure 8: Confusion matrices Tweets regression.



CAPTIONS: VALENCE

PREDICTED

		<.2	.2-.4	.4-.6	.6-.8	>.8
	<.2	3	40	23	9	0
T	.2-.4	3	92	119	45	9
R	.4-.6	1	27	158	88	12
U	.6-.8	0	7	57	134	69
E	>.8	0	0	7	35	62

CAPTIONS: AROUSAL

PREDICTED

		<.2	.2-.4	.4-.6	.6-.8	>.8
	<.2	9	29	22	2	0
T	.2-.4	2	76	128	43	3
R	.4-.6	0	39	194	101	17
U	.6-.8	0	6	57	142	51
E	>.8	0	0	5	34	40

CAPTIONS: DOMINANCE

PREDICTED

		<.2	.2-.4	.4-.6	.6-.8	>.8
	<.2	0	14	16	6	0
T	.2-.4	0	52	144	61	2
R	.4-.6	0	15	141	220	18
U	.6-.8	0	2	49	191	33
E	>.8	0	0	3	28	5

Figure 9: Confusion matrices Captions regression.

5.2 Tweets versus captions

Although (cost-corrected) accuracy is lower for Captions than for Tweets (see Table 9), the confusion in the Captions classification task is less pronounced than in Tweets (Figures 6 and 7), as also reflected in the higher macro F1 for Captions. This is in line with the observation that the drop in performance in the cross-domain setting is more outspoken when training on Tweets and testing on Captions than the other way around. Also in the regression task, the correlation coefficient for Captions is somewhat higher. The Captions model thus seems more robust than the Tweets model, which is also in line with the higher IAA scores that were obtained for this domain (Section 3.3).

As reported in Section 4.3.2, there is no undisputed positive effect of training on data coming from two domains. This is an interesting insight, as one might assume that training on more data (the size of the training data is almost doubled) is always beneficial. It seems that there is a domain discrepancy which makes cross-domain transfer suboptimal.

Note that the multi-domain setting did have a positive effect on Caption classification and Tweets regression, but not on Tweets classification (at least not when evaluating accuracy) and Captions regression. When the multi-domain setting would only have been beneficial in the case of regression but not in the case of classification, we would have made the assumption that the lack of positive effect in classification would have come from the different distribution of categorical labels, because such a divergence in label distribution does not occur for the dimensional labels. However, because the multi-domain setting was not beneficial for Tweets regression either, this suggests that the lacking generalizability is not only due to divergent label distributions, but also to the nature of the texts themselves.

We therefore visualise the last hidden state of the [CLS]-token in the optimised RobBERT embeddings, which is used for classification, using t-SNE plots. T-SNE (short for t-distributed stochastic neighbour embedding) is a non-linear dimensionality reduction algorithm that can map high-dimensional data to two (or more) dimensions, and is therefore often used to visualise hidden states of deep learning models. We trained for 5 epochs on the full Tweets subset, applied this model to the complete EmotioNL corpus (Tweets and Captions) and plotted the hidden state of each instance's [CLS]-token in a two-dimensional space. The resulting t-SNE plot is shown in Figure 10 on the left, after which the same process was repeated by training on the full Captions subset, as shown in the plot on the right. In both plots the instances are visually differentiated by color (Tweets = blue and Captions = red).

In both plots we clearly observe a separation of the instances from the Tweets and Captions subset, which can explain the lower performance in the cross-domain and multi-domain setup. For cross-domain and multi-domain transfer to work, the distribution of features in both domains should overlap. Therefore, for future work, domain adaptation techniques should be investigated. One suggestion could be to incorporate adversarial training as a way of forcing the model to become domain-invariant. Concretely, an emotion classifier and a domain discriminator could be trained jointly. While the emotion classifier learns the task of emotion detection based on the source data, the domain classifier will learn to discriminate between the source and target data. The learning objective of the domain discriminator is then reversed by applying a gradient reversal function in order to learn domain-independent representations and introduce domain confusion into the model.

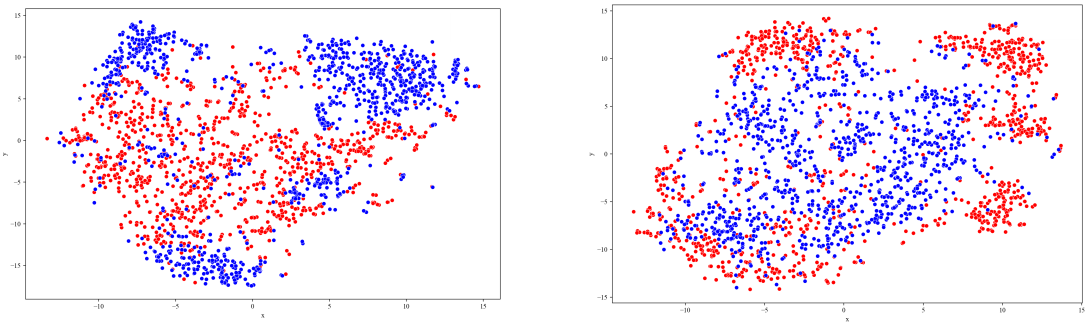


Figure 10: T-SNE plots of [CLS]-tokens after 5 epochs of finetuning RobBERT embeddings, trained on Tweets (left) and on Captions (right). ● = instances from Tweets; ● = instances from Captions.

5.3 Emotional components

We will now have a deeper look into the textual instances themselves and perform an error analysis in which we investigate how the way emotions are verbalised in text affects the performance of automatic emotion prediction. We therefore turn to the Component Process Model of emotion (CPM) by Scherer (2005).

According to the CPM, emotion is defined as an episode of changes in subsystems of an organism, triggered by an internal or external stimulus. These changes are related to five components of emotion, namely: 1) a subjective feelings component (the subjective emotional experience), 2) a cognitive component (appraisal of an event), 3) a motivational component (action tendencies), 4) a motor expression component (facial and vocal expression), and 5) a neurophysiological component (bodily symptoms) (Scherer 2005).

Casel et al. (2021) hypothesized that the way in which emotions are verbalised in text follows the component process model. Emotion can thus be expressed by describing a feeling (e.g. "I am sad"), reporting a cognitive appraisal (e.g. "It was a great party"), describing an action tendency (e.g. "I wanted to run away"), describing facial or vocal expressions (e.g. "laughing out loud"), or bodily changes (e.g. "my heart was pounding"). They annotated existing emotion corpora (1,000 instances from literature and 2,041 Tweets) with the emotion component classes (*subjective feelings*, *cognitive appraisal*, *motivational action tendencies*, *motor expressions* and *neurophysiological symptoms*). In the Twitter corpus, the *subjective feelings* and *cognitive appraisal* component were most prevalent (present in 75% and 32% of the tweets respectively), while in the literature corpus, the *cognitive appraisal* and *motor expressions* components were predominant (61% and 44%).

We now investigate how the emotion components are verbalised in the textual instances in EmotioNL, and whether there is a link between the components that are verbalised and the errors made by an emotion detection model. For example, instances that describe subjective feelings might be easier to classify than instances in which the emotion is solely verbalised by an action tendency.

Therefore, we took a subset of 10% of the data (i.e. 200 instances; same emotion label distribution as in the complete dataset) and checked which ones of the five above-mentioned component classes were present in each instance. From the 200 instances, 34 instances were neutral and were not analysed, as these naturally will not contain emotional components. Of the 166 remaining instances, 78 are tweets and 88 captions.

We used the guidelines from Casel et al. (2021) to assess the emotion components in the subset. Each textual instance could contain multiple component classes. An important addition to the guidelines though, is the judgment of emojis: when the emoji depicts a face or hand gesture, we see it as an instance of a *motor expression*, except when the emoji is irrelevant.

Table 11 shows the percentage of instances for which a certain emotion component was indicated as present. The *motivational action tendencies* and *neurophysiological symptoms* component were (almost) never expressed in either Tweets or Captions. Similar to the findings of Casel et al. (2021), the *cognitive appraisal* component appeared a lot in both domains (in 62% of the Tweets and 56% of Captions). However, while they found that the *subjective feelings* component often appeared in Tweets, this component was only predominant in the Captions subcorpus of EmotioNL (41% in Captions while only 12% in Tweets). In the Tweets subcorpus, the *motor expression* component was most prevalent (72%), mainly because of the presence of emojis.

We now look at the effect of the expressed emotional component on the prediction performance of our automatic emotion detection model (predictions from in-domain RobBERT). Figure 11 visualizes the number of correct and incorrect predictions for emotion classification, grouped by emotion component and domain. We observe that instances in which a *cognitive appraisal* is described, are more often correctly classified. *Motor expressions* are not necessarily easier to classify in Captions, but they are in Tweets, probably because of the use of emojis. Contrary to our expectations, *subjective feelings* are not easier to classify.

Finally, some instances were selected from the component-annotated subset to further investigate the relation between certain predictions and the underlying emotion components. These examples are listed in Table 12 (an English translation of the instances is given in Table 13 in the Appendix).

Instances 1–4 are examples in which *subjective feelings* are verbalised. The first two examples contain an explicit mention of emotional states ("I'm nervous" and "I'm really happy"). Unsurprisingly, these instances

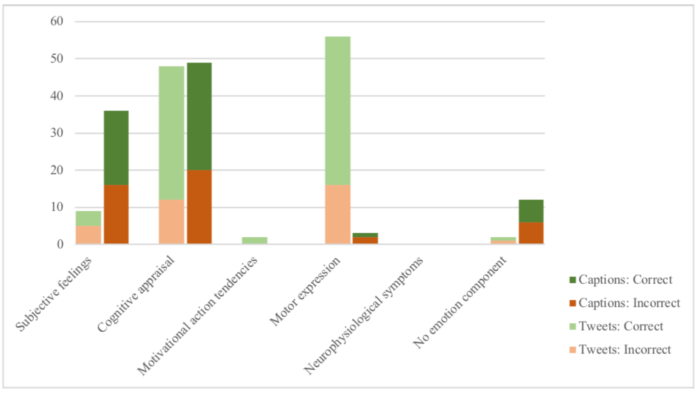


Figure 11: Correct and incorrect predictions for emotion classification, grouped by emotion component.

were correctly classified by the model. Although examples 3 and 4 are verbalisations of *subjective feelings* as well, the emotion is much less explicit here. These instances are wrongly classified, in example 3 probably because of the use of a kissing emoji (which made the classifier predict *joy*, while the true emotion was *fear*). For a human reader, example 4 is rather clearly expressing *sadness*, but this nuance is too implicit for the classifier.

Instances 5–8 all express *cognitive appraisals* (e.g. "pathetic and ridiculous", "it's ambiguous", "creepy shit", "I think I'm looking chic"). Indeed, these sentences seem rather straightforward to classify. However, sentences 7 and 8 are incorrectly classified, but actually, the labels predicted by the system are not completely illogical. Some people might judge example 7 as *anger* as well, and example 8 is in fact rather mild in emotional intensity, which could lead to judging it as *neutral*.

There were only two instances of *motivational action tendencies* in the component-annotated subset, of which instance 9 is an example. Still, it is questionable whether everyone would agree that the threat expressed in this message actually represents an action tendency.

Motor expressions are expressed in instances 10–13: in 10 and 13 by means of an emoji, in 11 by means of a description ("my mouth fell open") and an emoji, and in 12 by an exclamation ("buuuh"). While the description and exclamation do not really help the classification, emojis usually do. However, in example 13, we are dealing with irony, in which case the emoji is actually confusing the classifier.

In instances 14–16, no emotional components were expressed. This makes classification hard (even for humans), as the emotion is conveyed very implicitly (e.g. "the sun is shining", "this makes me think", "she stole my money"), and we really should trust on world knowledge.

This error analysis reveals that instances verbalising the *cognitive appraisal* component of emotion are easier to classify, as these are usually rather explicit. Expressions of the *subjective feelings* component are

	Subjective feelings	**Cognitive appraisal**	**Motivational action tendencies**	**Motor expression**	**Neuro-physiological symptoms**	**No emotion components**
Tweets	12% (9)	62% (48)	3% (2)	72% (56)	0%	3% (2)
Captions	41% (36)	56% (49)	0% (0)	3% (3)	0%	14% (12)
All	*27% (45)*	*58% (97)*	*2% (2)*	*39% (65)*	*0%*	*5% (9)*

Table 11: Percentage (and absolute number) of verbalised emotion components in the annotated subsets (78 tweets and 88 captions).

	Instance	Gold	Pred.	Components
1	Kziet nie meer. Kzin te nerveus. Kzin echt nerveus.	fear	fear	Subjective feelings
2	Ik voel mij echt gelukkig nu op dit moment.	joy	joy	Subjective feelings
3	@BiancavDijcke Heel apart he 😚 Aan dat idee moet ik best nog wel wennen. Dat mijn verleden nu ineens gedeeld gaat worden.... 😚	fear	joy	Subjective feelings
4	Het komt nu gewoon efkes allemaal heel hard binnen.	sadness	anger	Subjective feelings
5	Verwijder me dan maar van de prive insta zodat ik jullie fototjes niet zie 🙄 zielig en belachelijk is dat sorry	anger	anger	Cognitive appraisal, motor expression (emoji)
6	Eigenlijk dubbelzinnig weeral. Ik zie de maatschappij zo neig veranderen. Jullie zien het ook allemaal veranderen, ma gulle loopt erin mee.	sadness	sadness	Cognitive appraisal
7	Over #fakenews gesproken, als dit ooit als app released gaat worden 😳 Creepy shit, stemmen nabootsen met max 1 min aan bronsmateriaal	fear	anger	Cognitive appraisal, motor expression (emoji)
8	Ik vind mijn eigen der echt wel chic uitzien nu.	joy	neutral	Cognitive appraisal
9	@luzvanmaele Hahahah ni zo bedoeld 😤 maar beter probeer je me ni kwaad te maken of tloopt ni goe af luz puss	anger	anger	Action tendency
10	Omg de frituutpan staat alweer aan 😂 zit weer aan de patat 😅 😅	joy	joy	Motor expression (emoji)
11	Echt gebeurd alleen ik rende niet zo hard mn mond klapte open tot op de grond 😦 wtf.. Na 4 jr 😳	fear	sadness	Motor expression
12	Allee zie. Buuuh.	anger	fear	Motor expression
13	@geertwilderspvv Ik vind het wel grappig van je 😂 zelf niet mee doen aan debatten maar van buiten af het debat trollen via twitter.	anger	joy	Cognitive appraisal, motor expression (emoji)
14	Çavakes, Jef? Tzonneke schijnt eh.	joy	joy	/
15	@barteradus_bart @volkskrant Ik vind mijzelf geen rascist, maar dit artikel triggert mij wel tot nadenken 😉 daar word niemand slechter van	joy	anger	/
16	Ja en ik zeg juist zij heeft mij mn geld gepikt bij mij thuis.	anger	neutral	/

Table 12: Selection of instances with their emotion component annotation, gold emotion label and predicted emotion label.

not necessarily easy to classify, unless the emotional states are explicitly mentioned. Further, we see that some ‘wrong’ decisions of the classifier are actually very understandable, as judging emotions might be very personal and even humans would often disagree. This again advocates for cost-corrected accuracy, which accounts to some extent for this variability between annotators and allows for a fairer evaluation of the model.

6. Conclusion

We presented EmotioNL, a new dataset for Dutch emotion detection with 1,000 Dutch Twitter messages and 1,000 captions of reality TV-shows, annotated with the emotion categories *anger*, *fear*, *joy*, *love*, *sadness* and *neutral*, and the emotional dimensions *valence*, *arousal* and *dominance* (VAD). We believe this dataset is a useful resource for improving the state of the art in Dutch emotion detection, as it covers two genres, various topics per genre and has a bi-representational format.

Furthermore, we introduced a new metric, cost-corrected accuracy, which takes into account the cost or severity of a misclassification, and thus allows for a fairer evaluation of emotion detection models.

We evaluated the state-of-the-art transformer models BERTje and RobBERT on this new dataset and found that RobBERT outperformed BERTje for most subtasks. When investigating the portability of the models across domains, we observed that models do not generalize well across domains, and that even training on both domains does not necessarily help performance. Domain adaption techniques might tackle this, which we leave to investigate for further research.

The paper ends with an extensive discussion and error analysis, which suggests that emotion classification is more challenging than VAD regression and that the models trained on Captions are more robust than the ones trained on Tweets, probably because of the higher inter-annotator agreement in the former. By annotating a subset of the data according to the Component Process Model, we tried to link emotion components to errors made by emotion detection models. The *cognitive appraisal* component was prevalent in both domains, accompanied by the *motor expression* and *subjective feelings component* in Tweets and Captions, respectively. We found that explicit mentions of emotional states in *subjective feelings* or instances verbalising the *cognitive appraisal* component are easier to classify, as they usually are rather explicit.

Overall, emotion detection remains a difficult task, as especially reflected in the low classification scores. Given that the results for VAD regression seem more promising, leveraging dimensional representations in aiding emotion classification might be an interesting angle for future research.

Acknowledgements

This work was funded by the Research Foundation–Flanders under a Strategic Basic Research fellowship with Grant No. 3S004019.

References

Ahmad, Zishan, Raghav Jindal, Asif Ekbal, and Pushpak Bhattachharyya (2020), Borrow from rich cousin: Transfer learning for emotion detection using cross lingual embedding, *Expert Systems with Applications* **139**, pp. 112851, Elsevier.

Alm, Cecilia Ovesdotter, Dan Roth, and Richard Sproat (2005), Emotions from text: Machine learning for text-based emotion prediction, *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Vancouver, Canada, pp. 579–586.

Amigó, Enrique, Julio Gonzalo, Stefano Mizzaro, and Jorge Carrillo-de Albornoz (2020), An effectiveness metric for ordinal classification: Formal properties and experimental results, *arXiv preprint arXiv:2006.01245*.

Boot, Peter, Hanna Zijlstra, and Rinie Geenen (2017), The dutch translation of the linguistic inquiry and word count (liwc) 2007 dictionary, *Dutch Journal of Applied Linguistics* **6** (1), pp. 65–76, John Benjamins.

Buechel, Sven and Udo Hahn (2016), Emotion analysis as a regression problem - Dimensional models and their implications on emotion representation and metrical evaluation, *ECAI 2016 - Proceedings of the 22nd European Conference on Artificial Intelligence*, IOS Press, The Hague, The Netherlands, pp. 1114–1122.

Buechel, Sven and Udo Hahn (2018), Word emotion induction for multiple languages as a deep multi-task learning problem, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Association for Computational Linguistics, New Orleans, Louisiana, pp. 1907–1918.

Casel, Felix, Amelie Heindl, and Roman Klinger (2021), Emotion recognition under consideration of the emotion Component Process Model, *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, KONVENS 2021 Organizers, Düsseldorf, Germany, pp. 49–61.

Chatterjee, Ankush, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal (2019), SemEval-2019 task 3: EmoContext contextual emotion detection in text, *Proceedings of the 13th International Workshop on Semantic Evaluation*, Association for Computational Linguistics, Minneapolis, Minnesota, USA, pp. 39–48.

Daumé III, Hal (2007), Frustratingly easy domain adaptation, *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Association for Computational Linguistics, Prague, Czech Republic, pp. 256–263.

Daume III, Hal and Daniel Marcu (2006), Domain adaptation for statistical classifiers, *Journal of artificial Intelligence research* **26**, pp. 101–126.

De Bruyne, Luna, Orphée De Clercq, and Veronique Hoste (2020), An emotional mess! Deciding on a framework for building a Dutch emotion-annotated corpus, *The International Conference on Language Resources and Evaluation 2020, LREC 2020*, European Language Resources Association (ELRA), pp. 1636–1644.

De Bruyne, Luna, Orphée De Clercq, and Véronique Hoste (2021a), Annotating affective dimensions in user-generated content, *Language Resources and Evaluation* pp. 1–29, Springer.

De Bruyne, Luna, Orphée De Clercq, and Véronique Hoste (2021b), Emotional RobBERT and insensitive BERTje: Combining transformers and affect lexica for Dutch emotion detection, *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp. 257–263.

De Clercq, Orphée, Els Lefever, Gilles Jacobs, Tijl Carpels, and Véronique Hoste (2017), Towards an integrated pipeline for aspect-based sentiment analysis in various domains, *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Association for Computational Linguistics, Copenhagen, Denmark, pp. 136–142.

De Smedt, Tom and Walter Daelemans (2012), "Vreselijk mooi!" (terribly beautiful): A subjectivity lexicon for Dutch adjectives., *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, European Language Resources Association (ELRA), Istanbul, Turkey, pp. 3568–3572.

de Vries, Wietse, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim (2019), Bertje: A Dutch BERT model, *arXiv preprint arXiv:1912.09582*.

Delobelle, Pieter, Thomas Winters, and Bettina Berendt (2020), RobBERT: A Dutch RoBERTa-based language model, *arXiv preprint arXiv:2001.06286*.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019), BERT: Pre-training of deep bidirectional transformers for language understanding, *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)*, Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4171–4186.

Du, Chunning, Haifeng Sun, Jingyu Wang, Qi Qi, and Jianxin Liao (2020), Adversarial and domain-aware bert for cross-domain sentiment analysis, *Proceedings of the 58th annual meeting of the Association for Computational Linguistics*, pp. 4019–4028.

George, Nysia I, Tzu-Pin Lu, and Ching-Wei Chang (2016), Cost-sensitive performance metric for comparing multiple ordinal classifiers, *Artificial intelligence research* **5** (1), pp. 135, NIH Public Access.

Graterol, Wilfredo, Jose Diaz-Amado, Yudith Cardinale, Irvin Dongo, Edmundo Lopes-Silva, and Cleia Santos-Libarino (2021), Emotion detection for social robots based on NLP transformers and an emotion ontology, *Sensors* **21** (4), pp. 1322, Multidisciplinary Digital Publishing Institute.

Holzman, Lars E and William M Pottenger (2003), Classification of emotions in internet chat: An application of machine learning using speech phonemes, *Retrieved November* **27** (2011), pp. 50, Citeseer.

Jijkoun, Valentin and Katja Hofmann (2009), Generating a non-English subjectivity lexicon: Relations that matter, *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pp. 398–405.

Kiritchenko, Svetlana and Saif Mohammad (2017), Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 465–470.

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

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

Ljubešić, Nikola, Ilia Markov, Darja Fišer, and Walter Daelemans (2020), The lilah emotion lexicon of croatian, dutch and slovene, *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pp. 153–157.

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

Metzler, Hannah, Bernard Rimé, Max Pellert, Thomas Niederkrotenthaler, Anna Di Natale, and David Garcia (2021), Collective emotions during the covid-19 outbreak, PsyArXiv.

Mohammad, Saif, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko (2018), SemEval-2018 task 1: Affect in tweets, *Proceedings of The 12th International Workshop on Semantic Evaluation*, Association for Computational Linguistics, New Orleans, Louisiana, pp. 1–17.

Öhman, Emily, Kaisla Kajava, Jörg Tiedemann, and Timo Honkela (2018), Creating a dataset for multilingual fine-grained emotion-detection using gamification-based annotation, *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Association for Computational Linguistics, Brussels, Belgium, pp. 24–30.

Oostdijk, Nelleke, Martin Reynaert, Véronique Hoste, and Ineke Schuurman (2013), The construction of a 500-million-word reference corpus of contemporary written Dutch, *Essential speech and language technology for Dutch*, Springer, Berlin, Heidelberg, pp. 219–247.

Ordelman, Roeland, Franciska de Jong, Arjan Van Hessen, and Hendri Hondorp (2007), TwNC: a multifaceted Dutch news corpus, *ELRA Newsletter* **12** (3/4), pp. 4–7.

Plutchik, Robert (1980), A general psychoevolutionary theory of emotion, *in* Plutchik, Robert and Henry Kellerman, editors, *Theories of Emotion*, Academic Press, New York, pp. 3–33.

Rescorla, Robert A, Allan R Wagner, et al. (1972), A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement, *Classical conditioning II: Current research and theory* **2**, pp. 64–99, New-York.

Roumen, YN (2021), *Examining gender differences in language: A computational analysis of emotion in the dutch veteran institute oral history archive*, B.S. thesis.

Scherer, Klaus R (2005), What are emotions? And how can they be measured?, *Social science information* **44** (4), pp. 695–729, Sage Publications Sage CA: Thousand Oaks, CA.

Suárez, Pedro Javier Ortiz, Benoît Sagot, and Laurent Romary (2019), Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures, *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*, Leibniz-Institut für Deutsche Sprache.

Tromp, Erik and Mykola Pechenizkiy (2014), Rule-based emotion detection on social media: putting tweets on plutchik's wheel, *arXiv preprint arXiv:1412.4682*.

Vaassen, Frederik (2014), *Measuring emotion*, PhD thesis, Ph. D. dissertation, Dept. Taalkunde, Universiteit Antwerpen, Antwerp, Belgium.

Vaassen, Frederik and Walter Daelemans (2011), Automatic emotion classification for interpersonal communication, *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011)*, pp. 104–110.

Vaassen, Frederik, Jeroen Wauters, Frederik Van Broeckhoven, Maarten Van Overveldt, Walter Daelemans, and Koen Eneman (2012), delearyous: Training interpersonal communication skills using unconstrained text input, *Proc. of ECGBL* pp. 505–513.

Van de Kauter, Marjan, Diane Breesch, and Veronique Hoste (2015), Fine-grained analysis of explicit and implicit sentiment in financial news articles, *Expert Systems with Applications* **42** (11), pp. 4999–5010, Elsevier.

Van der Burgh, Benjamin and Suzan Verberne (2019), The merits of universal language model fine-tuning for small datasets–A case with Dutch book reviews, *arXiv preprint arXiv:1910.00896*.

van der Zwaan, Janneke M, Inger Leemans, Erika Kuijpers, and Isa Maks (2015), Heem, a complex model for mining emotions in historical text, *2015 IEEE 11th International Conference on e-Science*, IEEE, pp. 22–30.

van Lange, Milan and Ralf Futselaar (2021), Vehemence and victims: Emotion mining historical parliamentary debates on war victims in the netherlands, *DH Benelux Journal* **3**, pp. 61–79.

Waegeman, Willem, Bernard De Baets, and Luc Boullart (2006), A comparison of different ROC measures for ordinal regression, *Proceedings of the CML 2006 workshop on ROC Analysis in Machine Learning*, Citeseer.

Wang, Shihan, Marijn Schraagen, Erik Tjong Kim Sang, and Mehdi Dastani (2020), Dutch general public reaction on governmental COVID-19 measures and announcements in Twitter data, *arXiv preprint arXiv:2006.07283*.

Zhang, Yuhong, Xuegang Hu, Peipei Li, Lei Li, and Xindong Wu (2015), Cross-domain sentiment classification-feature divergence, polarity divergence or both?, *Pattern recognition letters* **65**, pp. 44–50, Elsevier.

Appendix A. EmotioNL Annotation Guidelines

These guidelines are formulated within the scope of the creation of a Dutch emotion corpus. A set of 1,000 Dutch tweets and 1,000 captions will be annotated for emotions, viewed from the perspective of the author/utterer. You will label the instances with a set of 5 emotions, namely *anger*, *fear*, *joy*, *love* and *sadness*, or the category *neutral*, and with the emotional dimensions *valence*, *arousal* and *dominance*. For the latter, you will use the best-worst scaling approach.

The texts that you will annotate are Dutch tweets generated by various Twitter users and captions from reality TV-shows. The goal is to reveal the emotional state of the author when the tweet was written or the speaker when the utterance was expressed. Therefore, you will project yourself into the perspective of the writer/speaker and imagine what the author/speaker must have felt when writing the tweet/expressing the utterance. You will see a text instance, and then have to check the box of the corresponding category (*anger*, *fear*, *joy*, *love*, *sadness* or *neutral*). Only indicate the most appropriate emotion. For the best-worst scaling part, you will see three blocks of four different sentences. The sentences are the same for the three blocks. In the first block, you will indicate the sentences with the highest *valence* and lowest *valence*, in the second block the sentences with the highest and lowest *arousal*, and in the third one you indicate the sentences with the highest and lowest *dominance*. More information on the dimensions *valence*, *arousal* and *dominance* and the concept best-worst scaling is given further in these guidelines.

A.1 Neutral or emotional

First, determine whether the instance is subjective or not. For neutral instances or objective instances (instances without any emotion), you indicate the *neutral* category.

A.1.1 Pitfalls

The most common pitfall is to let oneself be guided by emojis too much. Using an emoji does not necessarily mean that there is an emotion conveyed. Sometimes, emojis are used to make a message less ‘dry’, without really wanting to express an emotion. Such instances can still be neutral, although there is an emoji present (see Example 1). Emojis are also often used in advertisements. The authors of such advertisements want to evoke an emotion in the reader, rather than to express an emotion themselves. Such advertisements should be classified as objective (see Example 2). Also misleading trigger words can be a pitfall. One could be tempted to annotate Example 3 with the label *anger*. However, the writer of this tweet probably was not feeling angry when he wrote this comment, and the tweet can rather be seen as a neutral statement.

(1) what do you mean 😊 😊 😊

(2) We’re giving away 5 lucky people a free pair of #Gate socks 🤗 !!

(3) #Anger or #wrath is an intense emotional response.

A.2 Emotional categories

If you have not indicated the instance as *neutral*, you need to indicate the emotion that was felt by the author/speaker. You can choose between *anger*, *fear*, *joy*, *love* and *sadness*. Only indicate the most appropriate emotion.

A.2.1 Pitfalls

The goal is not to recognize emotions solely by means of textual clues (although this possibly may help), but also to track implicit emotions. The context is extremely important, and we emphasize that you should not rely on trigger words or emojis only. Here follow some examples of common pitfalls:

(4) Danish pastries...oh know my nightmare! Old boots is a good description #GBBO

(5) omg Im really excited for the new series ❤️ ❤️ ❤️ ❤️ ❤️ ❤️

(6) I feel so blessed to get ocular migraines.

(7) iPhone 6 battery drains so fast since last update and shuts down at 40%.

In Example 4, the word nightmare is used, but this trigger word is misleading: the expression has essentially nothing to do with *fear*, but rather with *disgust*, *irritation* or *anger*. In Example 5, the heart emojis suggest that the emotion *love* is expressed, but this is rather an instance of *enthusiasm* or *joy*. A good example of irony can be found in Example 6, in which the actual meaning of an utterance is different from what is literally enunciated. Instead of *contentment* or *joy*, the writer of this tweet actually wanted to express *frustration* or *anger*. The last example does not contain any emotional trigger words, but through world knowledge we know that the writer of this sentence probably wanted to express *irritation* or *anger*. This is an example of implicit sentiment. Again, it is the context that is the most important for this annotation task, not the individual textual clues or trigger words.

A.2.2 Examples

Here we will discuss the five emotion labels *anger*, *fear*, *joy*, *love* and *sadness* a bit more thoroughly. For every category, we give some related terms and example sentences.

Anger
Related terms: *rage, outrage, fury, wrath, hostility, ferocity, bitterness, hate, loathing, scorn, spite, vengefulness, dislike, resentment, disappointment, dismay, displeasure, disgust, revulsion, contempt, frustration, exasperation, irritation, aggravation, agitation, annoyance, grouchiness, grumpiness*

Examples:

(8) WHY TF DO BROKE BOYS KEEP TRYNA FIND GF's? GET UR FUCKIN $$$ RIGHT B4 I SMACK YO BITCHASS, she deserves 2be happy & u don't deserve that ass

(9) Danish pastries...oh know my nightmare! Old boots is a good description #GBBO

(10) Cheque book request - declined, Advance against salary - declined, loan installment deferment- declined. Feedback received as banks discretion.. 👌🏻 👌🏻 👌🏻 Anymore benefits of being a Rak bank customer?? @RAKBANKlive @RAKBANKhelp #poorservice

(11) I don't understand people who come to the gym just to stand around and talk.

Fear
Related terms: *alarm, shock, fright, horror, terror, panic, hysteria, mortification, nervousness, anxiety, tenseness, uneasiness, apprehension, worry, distress, dread*

Examples:

(12) Second day in a row that I hear a knock on the window and no one. #haunted #scared #homealone

(13) Just a thought of all the upcoming expenses & bills brings me shivers. #scared #savemoney 😯 😯 😯

Joy
Related terms: *amusement, bliss, cheerfulness, gaiety, glee, jolliness, joviality, delight, enjoyment, gladness, happiness, jubilation, elation, satisfaction, ecstasy, euphoria, contentment, pleasure, enthusiasm, zeal, zest, excitement, thrill, exhilaration, optimism, eagerness, hope, pride, triumph*

Examples:

(14) ATTENTION EVERYONE YA GIRL PASSED HER SENIOR EXIT EXAM AND ACED THAT BITCH!!!! OFFICIALLY FINNA GRADUATE INNISHOEEEEEE IM SO EXCITED OMGGGGGGG 🙌🏾 🙌🏾 🙌🏾 🙌🏾 🙌🏾

(15) Beautiful day 😎

(16) My interview went well today, I can't wait to find out what happens. #excited #interview #jobinterview

(17) If my luck the rest of Fall goes anything like today, I think I'm going to like this season. #bestdayever #magic #work #snap

(18) i really really luv our fandom we have been fighting for 13 days how amazing how powerful we are

Love
Related terms: *adoration, affection, fondness, attraction, caring, tenderness, compassion, sentimentality, enthrallment, rapture, longing, lust, arousal, desire, passion, infatuation*

This relates to some kind of 'romantic love' or 'world love', and not to just 'liking' something.

Examples:

(19) @Za_buhmaid Happy birthday sweety .. sweet 21 hun hope u have a wonderful day and a wondrful joyful year better than the last one, Luv U ❤️

(20) @TheMandyMoore You are beyond wonderful. Your singing prowess is phenomenal but damn... I'm just elated to watch you act again. #ThisIsUs

(21) All I need is a day with you here by my side

(22) There is literally not a single male specimen on this entire planet that is hotter than Luke Evans. This is fact. Cannot be proven wrong even if you tried. 🔥 💙 🔥 🤤 He's so hot I can't even see straight.

Sadness
Related terms: *depression, despair, hopelessness, gloom, glumness, unhappiness, grief, sorrow, woe, misery, melancholy, pity, sympathy*

Examples:

(23) im having the worst week ever and i cant even go home yet to just sulk in my bed

(24) Today something terrible happened in our town. In our school. I'm so sorry for everyone's loss. My heart goes out to everyone affected. If anyone ever needs to talk, or is feeling sad/depressed, please dont hesitate to talk to me, I'm always here for whoever needs me, at any time

A.3 Emotional dimensions

You will give annotations for the emotional dimensions *valence*, *arousal* and *dominance*. We give the definition and some examples of these dimensions, as described by Mehrabian (1996).

A.3.1 Definitions and examples

Valence (also called pleasure or evaluation): a continuum ranging from extreme pain or unhappiness to extreme happiness or ecstasy (e.g. *cruelty*, *humiliation*, *disinterest* and *boredom* versus *excitement*, *relaxation*, *love* and *tranquility*).
TIP: try to capture the overall feeling of the utterer
1 – very negative
2 – rather negative
3 – neutral
4 – rather positive
5 – very positive

Arousal: denotes the level of mental alertness and physical activity or energy (e.g. *sleep*, *inactivity*, *boredom* and *relaxation* at the lower end versus *wakefulness*, *bodily tension*, *strenuous exercise*, and *concentration* at the higher end). The *arousal* dimension is not equal to emotion intensity.
TIP: try to imagine how the utterer would have said it

1 – very calm, monotonous, bored
2 – rather calm
3 – neutral
4 – faster, a bit excited
5 – fast, loud, dynamic voice use, gestures

Dominance: relates to social position; the feeling of control and influence over one's surroundings versus the feeling of being controlled or influenced by situations and others (e.g. *relaxation*, *power*, and *boldness* versus *anxiety*, *infatuation*, *fear*, and *loneliness*). Note that this applies to the feeling you have after a situation emerged. After all, not that many situations occur in which you are in complete control. However, it is possible to have a feeling of dominance after something happened that was out of your control (e.g. when you're yelling out of anger about a stolen phone). Likewise, it is possible to feel submissive after a similar event (e.g. when you're crying because the phone really meant a lot to you).
TIP: try to visualize the pose of the utterer

1 – cringed, small, feeling of being overpowered, very insecure
2 – introvert posture, rather insecure
3 – neutral
4 – more open posture, rather self-confident
5 – standing up, feeling of self-confidence and power

A.3.2 Best-worst scaling

Instead of scoring the dimensions *valence*, *arousal* and *dominance* on a 5-point scale, best-worst scaling will be used. You will see three times (one for each dimension) the same four instances. For each dimension, you will need to indicate the sentence with the highest score (best) and the sentence with the lowest score (worst): first for *valence*, then for *arousal* and lastly for *dominance*.

Treat objective sentences as neutral. In practice, this means that the chance is lower of indicating them as either 'best' or 'worst', so that they will be ignored.

Appendix B. Translations

	Instance
1	I don’t see it anymore. I’m too nervous. I’m really nervous.
2	I feel really happy at the moment.
3	@BiancavDijcke Very peculiar 😚 I still need to get used to that idea. That my past all of a sudden is gonna be shared.... 😚
4	It’s just really starting to hit me now.
5	Then just remove me from the private insta so that I can’t see your stupid pictures anymore 🙄 it is pathetic and ridiculous sorry
6	Actually ambiguous again. I see society changes so heavily. You see it changing too, but you just go with the flow.
7	Talking about #fakenews, if this is ever going to be released as an app 😳 Creepy shit, imitating voices with max 1 minute of source material
8	I think I’m looking chic right now.
9	@luzvanmaele Hahahah not meant like that 😤 but better to not make me angry or it’s not going to end well luz puss
10	Omg the frying pan is in use again 😂 eating fries again 😅 😅
11	True story only I didn’t run that hard my mouth fell open 😯 wtf.. After 4 yrs 😳
12	Now look. Buuuh.
13	@geertwilderspvv I think you’re funny 😂 not participating in debates but trolling the debate from outside via twitter.
14	What’s up, Jef? The sun is shining.
15	@barteradus_bart @volkskrant I don’t consider myself racist, but this article makes me think 😉 that doesn’t hurt anyone
16	Yes and I’m telling you she stole my money at my place.

Table 13: English translation of the selection of instances from Table 12.