Emotional RobBERT and Insensitive BERTje: Combining Transformers and Affect Lexica for Dutch Emotion Detection

Luna De Bruyne, Orphée De Clercq and Véronique Hoste
LT^3, Language and Translation Technology Team
Ghent University
{luna.debruyne, orphee.declercq, veronique.hoste}@ugent.be

Abstract

In a first step towards improving Dutch emotion detection, we try to combine the Dutch transformer models BERTje and RobBERT with lexicon-based methods. We propose two architectures: one in which lexicon information is directly injected into the transformer model and a meta-learning approach where predictions from transformers are combined with lexicon features. The models are tested on 1,000 Dutch tweets and 1,000 captions from TV-shows which have been manually annotated with emotion categories and dimensions. We find that RobBERT clearly outperforms BERTje, but that directly adding lexicon information to transformers does not improve performance. In the meta-learning approach, lexicon information does have a positive effect on BERTje, but not on RobBERT. This suggests that more emotional information is already contained within this latter language model.

1 Introduction

Computational analysis of affect in Dutch texts is mostly restricted to polarity analysis (negative/positive/neutral), for which we know a tradition of lexicon-based approaches. Recently, a BERT-based model, BERTje (de Vries et al., 2019), and a RoBERTa-based model, RobBERT (Delobelle et al., 2020), have been created for Dutch, and they have achieved promising results on the task of sentiment analysis. For emotion detection, however, these models have not yet been evaluated.

In a first step towards improving emotion detection for Dutch, we will evaluate BERTje and RobBERT on the task of emotion detection. Instead of casting aside the many efforts that have been made in the creation of Dutch sentiment and emotion lexica, we will investigate whether transformers and affect lexica can complement each other. Attempts of combining BERT models with additional features have already been successful for tasks like abusive language and sarcasm detection (Koufakou et al., 2020; Kumar et al., 2021).

We consider two architectures. In the first one, we inject lexicon information in the transformer model before the prediction layer. We do this by concatenating the `[CLS]` token of the target sentence (which BERT and RoBERTa models use as input for prediction) with a lexicon vector obtained from seven Dutch affect lexica. In the second approach, we employ a meta-learning architecture and use a support vector machine (SVM) that learns from the transformer model's output. The predictions from the transformer are concatenated with the lexicon vector and used as input for the SVM.

We evaluate our models on 1,000 Dutch Tweets and 1,000 transcribed utterances from Flemish TV-shows. As multiple researchers have emphasized the need of studying emotions not only in terms of basic emotions, but based on dimensions like valence, arousal and dominance as well (Buechel and Hahn, 2016; Mohammad and Kiritchenko, 2018), the data has been annotated in a bi-representational design: both with categorical annotations for anger, joy, fear, love, sadness or neutral, and scores for the dimensions valence, arousal and dominance.

First we will discuss related work on Dutch emotion detection and other attempts on combining transformer models with additional features in Section 2. In Section 3, we will describe the methodology of our experiments and in Section 4, we report the results. We end with a conclusion in Section 5.

2 Related Work

Although most emotion detection research deals with the English language, recent studies have shown interest in other languages as well, e.g. the recent work of Ahmad et al. (2020) for Hindi or the

Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 257–263
April 19, 2021.

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

SemEval-2018 task on Affect in Tweets, for which in addition to English data, Arabic and Spanish datasets were released (Mohammad et al., 2018). Moreover, research on multilingual emotion detection (Buechel and Hahn, 2018; Öhman et al., 2018) and sentiment analysis (Lo et al., 2017; Vilares et al., 2018) is emerging.

Concerning the automatic modelling of affect in Dutch, the main focus still is on sentiment analysis instead of fine-grained emotions. Many studies have used sentiment lexica for this purpose (e.g. Van de Kauter et al., 2015; De Clercq et al., 2017; Wang et al., 2020). Recently, transformer models like BERTje (de Vries et al., 2019) and RobBERT (Delobelle et al., 2020) have been used to classify reviews from the Dutch Book Reviews dataset as either positive or negative, in which RobBERT came out as best (accuracy of 95% versus 93%).

The only publicly available dataset for Dutch emotion detection is deLearyous (Vaassen and Daelemans, 2011). It consists of 740 Dutch sentences from conversations, annotated according to Leary's Rose (Leary, 1957). Vaassen and Daelemans (2011) found that classifying sentences from deLearyous into the quadrants or octants of Leary's Rose was difficult for machine learners, most likely because of the sparseness of the data and low inter-annotator agreement. However, after this study, no further studies on Dutch emotion detection were published. This clearly shows the need of new data and methods for this task.

Some researchers already revealed that combining BERT models with handcrafted features can have a positive effect on performance, e.g. for the task of essay scoring (accuracy of 71% versus 80%) (Uto et al., 2020) and sarcasm detection (F1-score of 78% versus 80%) (Kumar et al., 2021). Lexicon features have been combined with BERT as well, e.g. for abusive language detection (Koufakou et al., 2020), for which the authors found an improvement on four out of six datasets compared to a plain BERT model (maximum improvement of 3%). In all of these studies, the handcrafted features or lexicon features were injected into the transformer architecture by concatenating them with the BERT representation before the prediction layer.

3 Method

3.1 Data

We collect data from two domains: Twitter posts (Tweets subcorpus) and utterances from reality TV-shows (Captions). For the Tweets subcorpus, a list of 72 emojis was used as query in the Dutch tweets database Twiqs.nl, with as search period 1-1-2017 to 31-12-2017. Based on this one-year datadump we sampled a random subset of 1,000 tweets, but made sure that no duplicates or non-Dutch tweets were present in the dataset.

For Captions, episodes of three emotionally loaded Flemish reality TV-shows (*Blind getrouwd*; *Bloed, zweet en luxeproblemen* and *Ooit vrij*) were used. Three episodes per show were transcribed using a literal transcription method (without correcting colloquial elements). 1,000 utterances (sentences or short sequences of sentences) were selected from these transcripts, based on a rough screening of emotional content and more or less equally distributed over the shows (335 instances from *Blind getrouwd*, 331 from *Bloed, zweet en luxeproblemen* and 334 from *Ooit vrij*).

All data was annotated with both categorical labels and dimensions. For the categorical annotation, the instances were labeled with one out of six labels: joy, love, anger, fear, sadness, or neutral. For annotating the dimensions valence, arousal and dominance, best-worst scaling was employed as this was shown to be a reliable annotation method (De Bruyne et al., 2021). Per subcorpus, the 1,000 instances 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,

	A	F	J	L	S	N
Tweets	188	51	400	44	98	219
Captions	185	97	331	42	185	160

Table 2: Number of instances in each emotion category per subdataset. A = anger, F = fear, J = joy, L = love, S = sadness, N = neutral.

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). See Table 1 for an annotated example of an instance in each domain.

Table 2 lists the number of instances per emotion category in each domain. For the valence, arousal and dominance annotations, the mean ranges between 0.46 and 0.52 for all dimensions in both subsets, the standard deviation ranges between 0.18 and 0.22, the minimum between 0.05 and 0.07 and the maximum between 0.96 and 0.97. For the experiments, both datasets were split in 800 instances for training, 100 for validating and 100 for testing (same splits for all models/tasks).

3.2 Lexicon information

We investigate whether lexicon information and transformer models can be complementary either by injecting lexicon information directly into the transformer architecture or by using a meta-learning approach in which predictions from transformer models are combined with lexicon features. Both models require the creation of a lexicon vector per target sentence.

For the creation of this vector, we combine seven existing open-source Dutch sentiment and emotion lexica, namely Pattern (De Smedt and Daelemans, 2012), Duoman (Jijkoun and Hofmann, 2009), LIWC (Boot et al., 2017), NRC Emotion (Mohammad and Turney, 2013), NRC VAD (Mohammad, 2018), Memolon (Buechel et al., 2020) and the VAD norms by Moors et al. (2013). For each word in the target sentence, lexicon values are obtained through a lookup in each affect lexicon. These values are then averaged over the words in the target sentence. The vector is 33-dimensional, as all lexica include values for multiple emotion categories or dimensions which add up to 33 in total. For lexica that do not have entries for any of the words in the sentence, the respective value in the lexicon vector is 0.

3.3 Transformer model

In this architecture, we inject the lexicon information into the transformer model while fine-tuning the model on the emotion detection tasks: emotion classification and emotion regression with the dimensions valence, arousal and dominance (VAD), and this in both domains. This injection occurs just before the prediction layer by concatenating the `[CLS]` token, which is normally used on its own as input for the classification, with the lexicon vector. This concatenated vector goes through a pre-classifier (linear layer with 2,048 nodes) and then to the prediction layer with Sigmoid activation function. The model architecture is shown in Figure 1.

Two Dutch transformer models are investigated: BERTje (de Vries et al., 2019), based on BERT by Devlin et al. (2019) and RobBERT (Delobelle et al., 2020), the Dutch version of the robustly optimized RoBERTa (Liu et al., 2019). RobBERT is trained on 39GB of common crawl data (Suárez et al., 2019), while BERTje is trained on only 12GB (including multiple genres).

Both models are implemented with HuggingFace's Transformers library (Wolf et al., 2019). We use AdamW optimizer (Loshchilov and Hutter, 2017) and the ReduceLROnPlateau learning rate scheduler with $l_r = 5e - 5$. The loss function is Binary Cross Entropy for the classification task and Mean Squared Error loss for regression. The maximum sequence length is 64 tokens, batch size is 16 for Tweets and 64 for Captions. We set dropout to 0.2 and use GELU as activation function in the implementation of Hendrycks and Gimpel (2016). The `[CLS]` token based on the concatenation of the last four layers of the model is used for prediction (and concatenated with the lexicon vector). The maximum number of epochs is set to 100 with a patience of 5 for early stopping.

3.4 Meta-learner

In the second approach, we use a support vector machine as meta-learner that learns from the predictions of a transformer model. We apply the transformer model on the training set and the test set to extract probabilities and predictions for sentences. For the classification task we use the probabilities of the emotion classes as features (6 features) and for the regression task the predicted scores (3 features). During training, the output on the training set is accompanied by lexicon features (33 features)

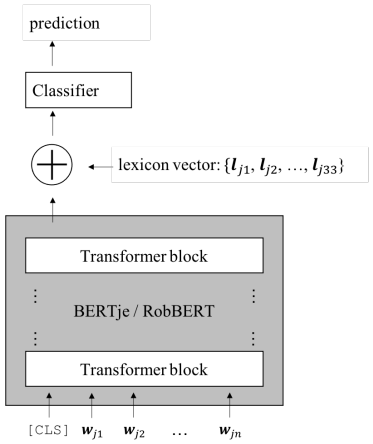

Figure 1: Transformer model with injection of lexicon features.

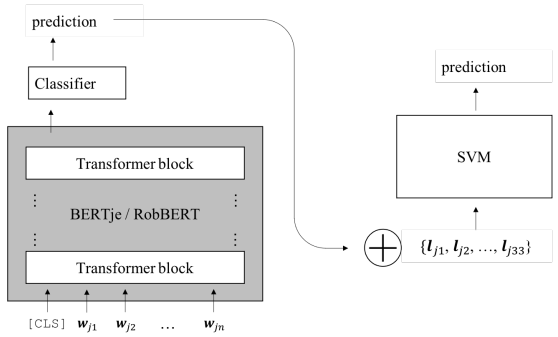

Figure 2: Meta-learner (SVM) with predictions from transformer and lexicon values as input features.

as input for the SVM. During testing, the output on the test set is combined with the lexicon features to feed the SVM. The model architecture is shown in Figure 2.

For all subtasks, we use a linear kernel and 1.0 as regularization parameter C. We use hinge loss and L2 penalty for classification and epsilon insensitive loss (L1 loss) for regression.

4 Results

4.1 Transformer model

We evaluate BERTje and RobBERT on the emotion classification (macro F1) and regression tasks (Pearson's r) for Tweets and Captions. We compare the results of the plain models and the models where the lexicon vector was injected.

Models are run ten times to account for variability and to be able to statistically compare the performance of different methods. The mean and standard deviation of these runs are reported in Table 3. Except for Captions classification, we observe that RobBERT is better for all datasets, with a notable margin: for Tweets classification and Captions regression there is an improvement of around 10% (from .16 to .26 and .60 to .68 respectively), and for Tweets regression the score almost doubled (from .36 to .70). This is in line with previous findings of RobBERT being more accurate in predicting sentiment (Delobelle et al., 2020), although our results are even more distinct. The robustly optimised RoBERTa framework and the larger corpus on which RobBERT was pre-trained clearly show their effect, although scores are in general still rather low. The results for the regression tasks seem more promising than for the classification task, although scores are difficult to compare as different metrics are used. We further notice that there is quite some variation between the runs, especially for the classification tasks.

Independent two-tailed t-tests were used for assessing whether the means of the models with and without lexicon vector are statistically different on a 5% significance level. For none of the datasets adding the lexicon vector was beneficial, as there was no significant difference in mean for any of the tasks (Tweets classification with BERTje: $t(9)$ = 1.3, p = 0.22; Tweets classification with RobBERT: $t(9)$ = 0.6, p = 0.54; Tweets regression with BERTje: $t(9)$ = 0.7, p = 0.50; Tweets regression with RobBERT: $t(9)$ = -0.6, p = 0.52; Captions classification with BERTje: $t(9)$ = 0.8, p = 0.42; Captions classification with RobBERT: $t(9)$ = -1.0, p = 0.31; Captions regression with BERTje: $t(9)$ = 0.5, p = 0.60; Captions regression with RobBERT: $t(9)$ = -0.6, p = 0.58). In contrast to previous studies regarding sarcasm and abusive language detection, we must thus conclude that this method of injecting lexicon information into the transformer model does not lead to higher performance on the task of emotion detection. We see two possible reasons for this: a) the lexica have no added value compared to the information that is already present in the language models, and/or b) the lexica do not have enough weight compared to the high-dimensional `CLS` token.

4.2 Meta-learner

In the meta-learning approach, we identify the best run (lowest loss on validation set) among the plain transformer models (i.e. without lexica), and apply that model on the training and test sets of the corresponding datasets. The probabilities (for classification) and predictions (for regression) are then

Transformers Model	Tweets F1 (SD)	Tweets r (SD)	Captions F1 (SD)	Captions r (SD)
plain BERTje	.16 (.05)	.36 (.03)	.26 (.04)	.60 (.02)
BERTje + lex	.14 (.02)	.35 (.03)	.24 (.04)	.59 (.02)
plain RobBERT	.26 (.09)	.69 (.03)	.22 (.08)	.68 (.02)
RobBERT + lex	.23 (.07)	.70 (.01)	.25 (.04)	.68 (.02)

Table 3: Results (macro F1-score and Pearson's r) of the plain transformer models (BERTje and RobBERT) versus the results with lexicon features added (BERTje/RobBERT + lex). F1-score and r is the average of 10 runs, standard deviation is shown between brackets.

Features	Tweets F1	Tweets r	Captions F1	Captions r
SVM lex	.16	.32	.12	.28
plain BERTje	.23	.38	**.28**	.61
SVM lex + BERTje	***.31***	***.41***	.26	***.62***
plain RobBERT	**.47**	.71	.30	.67
SVM lex + RobBERT	.42	.71	.30	***.68***

Table 4: Results (macro F1-score and Pearson's r) of the plain transformer model versus the meta-learner. The results of an SVM model with only lexicon features are given in the first line as reference. The best result is shown in bold or bold italics if the best model is the meta-learner approach.

fed into an SVM together with the lexicon features. The results of the original plain transformer models and meta-learning models are shown in Table 4.

We see that a meta-learner using the output of a transformer model combined with lexicon features outperforms the plain transformer model in the case of BERTje for three out of four tasks (all but Captions classification). In the case of RobBERT, the meta-learner outperforms the plain transformer model for only one dataset (Captions regression), but only to a minor extent. In contrast to the previous approach, where lexicon information was injected directly into the transformer model, the meta-learner does seem to be able to improve performance. Where the previous approach might have failed because of the lexica not having enough weight, the meta-learner seems to be better in exploiting the lexicon information, however only for BERTje. Possibly, emotions are more contained in the RobBERT language model because of the improved training framework and notably larger training corpus, making RobBERT a more emotional language model. Therefore, RobBERT benefits less from the added lexicon information in contrast to BERTje.

5 Conclusion

As the recently developed transformer models for Dutch, BERTje and RobBERT, have not yet been tested on emotion detection tasks, we evaluated them on 1,000 Dutch Tweets and 1,000 captions for an emotion classification and regression task. We found that RobBERT outperformed BERTje in almost all cases. Further, we investigated whether these models could be enhanced with lexicon features and proposed two methods: one in which a lexicon vector was concatenated with the transformer's `[CLS]` token before prediction, and a meta-learning approach. In the first method, adding lexicon information did not seem beneficial. For the second approach, however, we found that the meta-learner had a positive effect in half of the cases, and especially on the models relying on BERTje, whereas the meta-learner had almost no impact on RobBERT. This is probably because more emotional information is already contained within this language model.

Acknowledgements

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

References

Zishan Ahmad, 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*, 139:112851.

Peter Boot, 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):65–76.

Sven Buechel and Udo Hahn. 2016. Emotion analysis as a regression problem - dimensional models and their implications on emotion representation and metrical evaluation. In *ECAI*.

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

Sven Buechel, Susanna Rücker, and Udo Hahn. 2020. Learning and evaluating emotion lexicons for 91 languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1202–1217, Online. Association for Computational Linguistics.

Luna De Bruyne, Orphée De Clercq, and Véronique Hoste. 2021. Annotating affective dimensions in

user-generated content. *Language Resources and Evaluation*, pages 1–29.

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

Tom De Smedt and Walter Daelemans. 2012. " vreselijk mooi!"(terribly beautiful): A subjectivity lexicon for dutch adjectives. In *LREC*, pages 3568–3572.

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. Robbert: a dutch roberta-based language model. *arXiv preprint arXiv:2001.06286*.

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

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

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

Marjan Van de Kauter, Diane Breesch, and Veronique Hoste. 2015. Fine-grained analysis of explicit and implicit sentiment in financial news articles. *EXPERT SYSTEMS WITH APPLICATIONS*, 42(11):4999–5010.

Anna Koufakou, Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020. HurtBERT: Incorporating lexical features with BERT for the detection of abusive language. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 34–43, Online. Association for Computational Linguistics.

Avinash Kumar, Vishnu Teja Narapareddy, Pranjal Gupta, Veerubhotla Aditya Srikanth, Lalita Bhanu Murthy Neti, and Aruna Malapati. 2021. Adversarial and auxiliary features-aware bert for sarcasm detection. In *8th ACM IKDD CODS and 26th COMAD*, CODS COMAD 2021, page 163–170, New York, NY, USA. Association for Computing Machinery.

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

Yinhan Liu, 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*.

Siaw Ling Lo, Erik Cambria, Raymond Chiong, and David Cornforth. 2017. Multilingual sentiment analysis: from formal to informal and scarce resource languages. *Artificial Intelligence Review*, 48(4):499–527.

I Loshchilov and F Hutter. 2017. Fixing weight decay regularization. *arXiv preprint arXiv:1711.05101*.

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

Saif Mohammad and Svetlana Kiritchenko. 2018. Understanding emotions: A dataset of tweets to study interactions between affect categories. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Saif M. Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of The Annual Conference of the Association for Computational Linguistics (ACL)*, Melbourne, Australia.

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

Agnes Moors, Jan De Houwer, Dirk Hermans, Sabine Wanmaker, Kevin Van Schie, Anne-Laura Van Harmelen, Maarten De Schryver, Jeffrey De Winne, and Marc Brysbaert. 2013. Norms of valence, arousal, dominance, and age of acquisition for 4,300 dutch words. *Behavior research methods*, 45(1):169–177.

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

Robert A Rescorla, 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*, 2:64–99.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In *7th Workshop on the Challenges in the*

Management of Large Corpora (CMLC-7). Leibniz-Institut für Deutsche Sprache.

Masaki Uto, Yikuan Xie, and Maomi Ueno. 2020. Neural automated essay scoring incorporating handcrafted features. In *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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

David Vilares, Haiyun Peng, Ranjan Satapathy, and Erik Cambria. 2018. Babelsenticnet: a commonsense reasoning framework for multilingual sentiment analysis. In *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1292–1298. IEEE.

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

Shihan Wang, 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*.

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