Dimension is All You Need?
A Compact Dimensional Approach to Emotion Recognition

Matan Ben-Tov
Computer Science / Tel-Aviv University
bentov.matan@gmail.com

Shir Frenkel
Computer Science / Tel-Aviv University
frenkel.shir@gmail.com

Abstract

Precise extraction of specific emotions from text can further advance many fields - harmful behavior detection on the internet, psychological therapy, psychological research, digital assistants and also new datasets creation. There are different ways of approaching Emotion Recognition from text; In this paper we explore a creative yet simple architecture which utilizes a psychological dimensional emotion representation (called *VAD*[1]). Our proposed architecture uses VAD as an *intermediate* between the learned model and the emotions. For learning, comparing and evaluating the performance of our approach we use GoEmotions Dataset (Demszky et al., 2020), human-annotated English text-passages labeled with 28 different emotions. In this paper, we show that utilizing VAD in a compact BERT-based model[2] reaches comparable performance to the BERT baseline presented by the original paper (Demszky et al., 2020), we also discuss the advantages of such model.

Keywords: Emotion Recognition, GoEmotions, VAD

1 Introduction

While predicting the positiveness of a text (i.e. *Sentiment Analysis*) can be considered an "easy" problem, with simple models outperforming the human score on various datasets (e.g. *IMDB Reviews* dataset with 97% accuracy[3]), predicting a more fine-grained sentiment seems to pose a challenge to the same models. Intuitively, extracting specific emotions from text requires a deeper "understanding" of it, rather than the task of simple binary sentiment decision.

The common approach for fine-grained emotion classification task is, naturally, as a binary classification of *each* emotion. This approach treats emotions in a "binary fashion" (that is, we either recognize an emotion in a text or do not). This treatment might implicitly suggest that these emotions live in a *discrete* space. However, we know that emotions could be more complex than either "appear" or not. For example - an emotion can be present with specific degree of intensity. Moreover, from the recognition of one emotion, we might deduce some properties of another[4].

In psychology literature, asking these questions is practically pushing at an open door (2.3). A known method of representing emotions is the Valence-Arousal-Dominance (*VAD*) model (Russell and Mehrabian, 1977), which rates each emotion to 3-dimensional continuous space[5]: *Valence* (the degree of pleasantness or unpleasantness of an emotion), *Arousal* (degree of calmness or excitement), and *Dominance* (the degree of perceived control ranging from submissive to dominant). Formally: $VAD_{map} : emotion \mapsto [0,1]^3$. By rating 3 features of each emotion, VAD conveys a deeper meaning of the emotions as well as their correlations in different aspects. Thus, in our opinion, it seems to be a better fit for representing fine-grained emotions than discrete labeling.

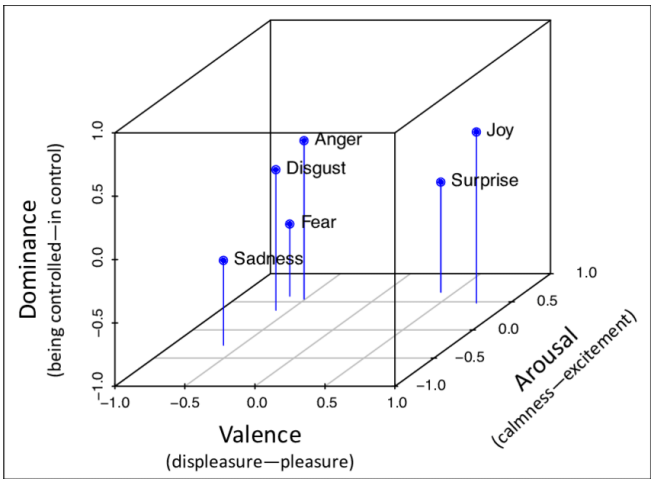

Figure 1: The 3D VAD Space

The stated above has motivated us to review the task for training fine-grained emotion detection; from the original *classification* model to our *VAD-based regression* model. Intuitively, one should expect the latter model

[1] Acronym of Valence-Arousal-Dominance, the 3 dimensions of representation.

[2] Our implementation in GitHub

[3] Leaderbords for IMDB-Reviews Dataset

[4] For example, although "Joy" and "Anger" might sound the opposite, both can be viewed as dominant and influencing emotions, yet can be distinguished by positivity (joy) and negativity (anger).

[5] An interactive VAD visualization

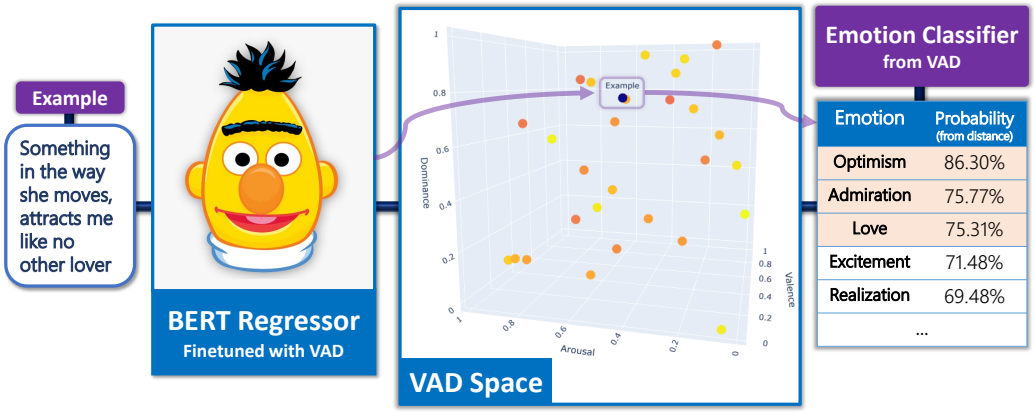

Figure 2: Our model's pipeline demoed with never-seen-before example.

to understand the relations between the emotions better, as it is being "tipped" by a well researched psychological representation. But an even stronger motivation for a VAD-based-model is the fact that, as opposed to the classification model in the baseline, this model will be able to classify never-seen-before emotions, as it learns the VAD space - rather than a predefined set of emotions. Further more, the model we'll present (3) can be later used both as a VAD-estimator model and as an emotion classifier, so we basically train one model and "get another one for free".

Our model is built of 2 major components: a BERT-based **regressor** (maps text to VAD space) and an **emotion-classifier** (maps VAD vectors back to labels). We train it on GoEmotions dataset (Demszky et al., 2020), after we mapped each label to VAD using (Mohammad, 2018) mapping. In our model, we first regress the text to VAD space, using BERT, then we use a much simple ML models to "classify-back" the emotion (visualized in Figure 2).

From the experiments we learn that a regression model followed by simple classifier can reach results **close** to the BERT baseline classifier.

However, despite the attempts we made and show in this paper, a simple VAD-based model does not seem to drastically outperform the strong baseline presented by ad-hoc classification with BERT. This result can be seen as another win for the deep-learning "doctrine", that claims that inserting human expertise could be harmful rather than helpful. Albeit, as we already claimed, as opposed to the ad-hoc baseline, our regression-based model is far more explainable[6], and can be reused on different datasets and with new labels.

In the following section (2) we present the data we use, and overview related work in both psychology and ML fields. In section 3 we thoroughly present our proposed model on its different variants followed by the experiments results (4). In section 5 we challenge our model with zero-shot experiments. Finally, in section 6, we conclude our discussion.

2 Related Work and Data

2.1 GoEmotions Dataset

GoEmotions dataset (Demszky et al., 2020) contain 58K English Reddit comments, multi-labeled for 27 emotion categories or Neutral. As of now, this dataset, as opposed to other datasets language-based emotion classification, is the larget and most fine-grained in terms of emotions. Each example in the dataset is assigned with at least one emotion. In practice 83% of the examples are labeled with precisely *one* emotion.

For our usage, we filtered out the multi-labeled examples[7] (that is, the remaining 17%), in order to focus our work on pure comparison of ad-hoc classification model against VAD-based model rather than involving unwanted variables such as VAD-mapping of multi-labels, etc.

GoEmotions' paper also presents a baseline which is BERT finetuned with a classification layer, achieving (only) 46% Macro-F1 score, showing the difficulty of fine-grained emotion classification.

2.2 VAD Dataset

NRC VAD Lexicon (Mohammad, 2018) is a crowd sourced lexicon, rating the valence, arousal and domi-

[6]For example, we can better understand our model's classification choices based on the VAD it outputs.

[7]Obviously, we re-trained the original paper's baseline on our dataset variation to present a fair comparison.

nance of 20K English words. That is, each example is a word mapped to a 3D vector in the VAD space.

For our end, we've extracted the VAD mapping of the 28 emotions in GoEmotions, and by that "labeled" each example of GoEmotions with VAD vector.

2.3 Approaches for Emotions Modeling

Despite the absence of consensus among psychological researchers regarding an emotion representation model, these models go way back and can be roughly divided to the following two groups:

- **Discrete / Categorical Models:** Often employ Ekman's six basic emotions (Ekman, 1992): Anger, Disgust, Fear, Joy, Sadness and Surprise. These emotions, by this paradigm, can unambiguously classify every emotional state.

- **Dimensional Models:** Often refer to the continuous 3-dimensional space of Valence-Arousal-Dominance (Russell and Mehrabian, 1977). This system embodies important information of each emotion by rating these 3 features, and allows measuring to distances between emotions. In addition, it is not limited to a finite set of labels (and there is no consensus on such set anyway), thus can be considered more expressive than the discrete model. Latest psychological studies proposes an even higher dimensional representation (Cowen et al., 2019).

In our model we will utilize the well-established representation model of VAD.

2.4 Works on GoEmotions

In the last couple of years there have been many works presenting fine-grained emotions detection ML models. In particular, many of these have attempted to challenge the baseline presented by GoEmotions. However, as far as we found out, none has succeeded improving *drastically*.

- **Combining dimensional and categorical emotions for a probabilistic NLP model** (Park et al., 2021): a RoBERTa-based probabilistic model combines VAD and categorical emotions labels to train an emotion-classifier. This model under-performs GoEmotions' baseline, but reaches comparable performance evaluated on other datasets.

- **Uncovering the Limits of Text-based Emotion Detection** (Alvarez-Gonzalez et al., 2021): This work shows a series of models experimented with fine-grained emotion detection (streches from BOW to BERT). Their best model[8] on GoEmotions reaches 48% Macro-F1. The paper also offers web interface for their model, which can help better understanding of the task we try to cope with.

- **Fine-Grained Emotion Prediction by Modeling Emotion Definitions** (Singh et al., 2021): Inspired by BERT's pre-training tasks, MLM (Masked Language Model) and NSP (Next Sentence Prediction), this paper presents analogical auxiliary tasks for finetuning on GoEmotions. For example, the "next sentence" in the new NSP task, is a list of emotions that can either fit or not to the original sentence. This creative approach outperform the baseline, reaching 52.3% Macro-F1 (currently GoEmotions' SOTA).

3 Model Architecture

Our model's structure (Figure 3) is composed of 3 primary parts:

- **VAD Mapper**: a static mapper from discrete emotions labels to VAD space, based on human annotations.

- **Regressor**: estimates a point in the VAD space that represent a given sentence.

- **Classifier**: given a point in the VAD space, it returns a emotion label.

In the following section we further describe each component and its variants.

3.1 VAD Mapper

The VAD mapping is part of the data processing, it is used for creating a dataset suitable for our regression model: we use this mapping on GoEmotions dataset, changing its targets from emotions labels to points in the VAD space. In addition, the classifier is trained on this mapping (as it embodies the reversed mapping). Our mappers are based on the NRC VAD lexicon (Mohammad, 2018), that maps English words to their VAD scores.

After exploring the NRC VAD mapping we discovered that the emotions distribution in the space is not uniform (shown in Figure 4). There are emotions that are very close to each other and there are large areas with no emotions at all. We suspected that such uneven distribution would make the task difficult for our model. In order to disperse the emotions more evenly in the space, we use *quantile transformation*[9] scaling. This scaling transforms each dimension to follow a uniform distribution, based on its cumulative distribution function. This transformation preserves the rank of the values along each dimension but distort the distances between the points. As we will see in section 4.3, by tuning one of the parameters of this method[10], we obtained a scaled VAD mapping that gave slightly better classification results than the original NRC VAD mapping.

[8] The model applies DNN over every encoded token of BERT and then reduces these vectors to the classification vector.

[9] sklearn quantile transform

[10] We tuned `n_quantiles` parameter, the larger it is, the more uniform the distribution is.

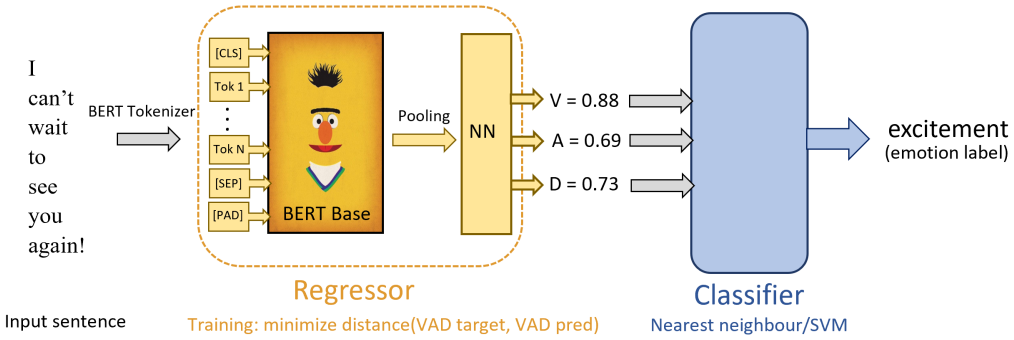

Figure 3: Overview of our model architecture.

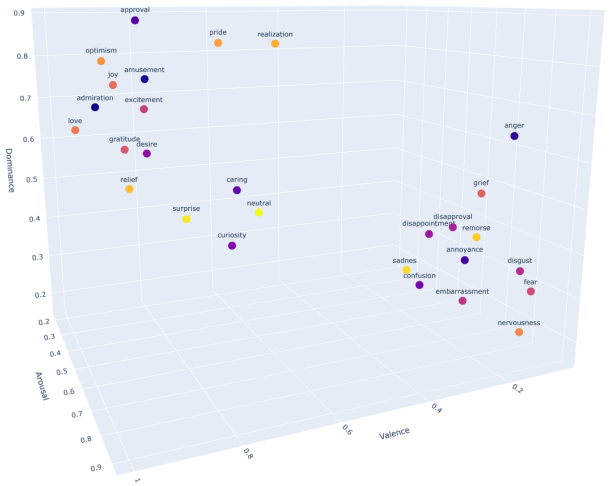

Figure 4: Emotions in the VAD space based on the NRC lexicon.

3.2 Regressor

Our regression model consists of two parts but trained as one piece. It is composed from a BERT base model and a small Neural Network that converts BERT's output (after pooling) to a 3D vector. Our goal is that this vector will be the corresponding VAD point for the given input.

We examine several loss functions for this regression model:

- **MAE:** Mean Absolute Error (L1 loss).
- **MSE:** Mean Squared Error (L2 loss).
- **MAE + CE:** Mean Squared Error combined with Cross Entropy (CE) Loss.

  This loss is an attempt to combine the regressor and the classifier together, to explore the possibility that a training with a loss that depends on the combination of the two stages together will lead to better results.

  The MAE + CE loss is a linear combination of MAE and CE loss, where the CE loss is based on the distances between the predicted point and the 28 VAD points corresponding to our 28 emotions. The formal definition of this loss is supplied in the appendix (7.1).

3.3 Classifier

The classifier converts back from a point in the VAD space to an emotion label. We use two types of classifiers:

- **Nearest Neighbour Classifier (1NN):** This classifier returns the emotion (out of 28 emotions) that is closest (in the VAD space) to the given VAD point. The distance is determined by a metric, such as the Euclidean metric or the Manhattan metric and the mapping from the 28 emotions to VAD is determined according to the *VAD mapper*.
- **SVM Classifier:** As the name suggests, this classifier is based on SVM - a supervised learning model, used in our case for classification analysis.

  We train the SVM model on our training set as follows: for each sentence s and label l from the training set, we run s through our regression model, getting a VAD point p. Then, we train the SVM on all (p, l) pairs.

  The SVM classifier returns a distribution over the 28 emotions labels. From this, we can easily get a mapping back to one emotion label, by choosing the most probable label.

4 Experiments

4.1 Experimental Details

In our experiments we use BERT-base-cased. We Fine-tune our model for 10 epochs, and present in this paper the results on the dataset, that were obtained by choosing the optimal checkpoint based on the dev set. The hyper parameters that were tuned: batch size, pooling, learning rate, noise addition (to the VAD target points), dropout, number of hidden layers in the Neural Network

(that converts BERT's output to VAD space). We also show here how different variation of each component in our model affect the results. The full details can be found on our implementation in GitHub[11], as well as how to reproduce the results.

Since our model is basically two models in one, we will first present the results of the regression model, which emits an evaluation of the given sentence in the VAD space, then we will present the results of the full model (which produces emotion tags) compared to GoEmotions baseline.

4.2 Experimenting the Regressor

In the following experiments, we test the quality of the regressor, composed from various VAD mappers and loss functions. As stated above, the purpose of the regressor is to emit a point in the VAD space that represents the emotion expressed in the given sentence (3.2).
For the regression task alone, we use two common loss functions to examine: MSE and MAE. We evaluate the performance of the regressor with the loss we are trying to minimize.

VAD mapper	Loss function	Loss value
VAD	MSE	0.037
VAD	MAE	0.116
VAD Scaled	MAE	0.127

Table 1: Regression to VAD space results. Each loss value corresponds to the given loss function.

The regression emits points in the VAD space: $[0, 1]^3$, hence, in the worst case, MSE loss will be $\sqrt{3}$ (≈ 0.173) and the MAE loss will be 3. From the above argument and the results in Table 1, it can be seen that our regression models achieve respectable results, which will serve us for GoEmotions classification task. Moreover, thanks to the generality of the VAD space versus emotion labeling, these regression models can be used for additional tasks, such as other emotion classification tasks, perhaps even with emotions our model has never seen, or simpler tasks like positive/negative sentiment analysis. We will explore some of those possibilities in section 5.
In addition, we can see that the original NRC VAD mapping gives us slightly better regression results.

4.3 Experimenting the Classifier

In this section we present the performances of our full classification model, which composed of a regressor (3.2) together with a classifier (3.3). We use accuracy and macro-F1 scores as evaluation metrics and compare our model to the baseline classification model, that was presented in GoEmotions paper, under the same conditions, which include the same training, dev and test sets.

From the results presented in Table 2, it can be seen that overall our model achieves comparable performance to the baseline model.

Model (mapper, loss, classifier)	Accuracy	Macro F1
GE baseline	0.557	**0.493**
VAD, MAE, 1NN	0.575	0.444
VAD, MAE, SVM	0.587	0.451
VAD, MSE, 1NN	0.295	0.117
VAD, MAE+CE, 1NN	0.589	0.435
VAD Scaled, MAE, 1NN	0.579	0.447
VAD Scaled, MAE, SVM	**0.59**	0.454

Table 2: Classification performances of our approach compared to GoEmotions baseline, on GoEmotions dataset.

Moreover, our model achieves better accuracy performance. On the other hand, according to the macro F1 metric, the baseline model outperforms our model. One of the reasons for the discrepancy between the performances over the above metrics is the fact that they average over different factors; accuracy averages over all examples, where macro-F1 metric is the average F1-score over the different classes (which in our case are the 28 emotions). Therefore, the accuracy metric is more affected by large classes (i.e. classes with many examples) compared to the macro-F1 metric. In our dataset, the distribution of emotions is imbalanced and causes such differences between the metrics.

Because of this bias of the accuracy metric and the imbalanced classes distribution, we believe that macro-F1 metric is more relevant for measuring our performance.

Our best performing model is *VAD Scaled-MAE-SVM*. In general, it can be seen that SVM classifier gave slightly better results than 1NN classifier and similarly, scaled VAD mapper did better than the original NRC VAD mapper, which explains why the above model gave the best performances (from our models).

As mentioned above, the scaled mapper achieved better results, although we saw in the regression experiments (4.2) that the regressor with the scaled VAD mapper gave worse performance. We presume that the scaled mapper is better than the original VAD mapper for the classification task because the scaled mapper maps the emotions to points that are more widely spread out in our three-dimensional space, making the translation back from VAD point to one emotion easier. On the other hand, it is important to remember that scaling changes distances between points, which causes some of the information that we get from the VAD representation to be lost, information that might be helpful for our model.

In addition, we can argue that SVM classifier yields better results because it is more tuned to the output of our regressor, in fact, it is fit to the regressor output. In contrast, the 1NN classifier is a static mapper, that is based on our VAD mapper and does not depend at all on the regressor outputs.

We can also notice that there is a significant gap between the performance of MAE-based models and MSE-based models: the model that uses MSE loss together

[11] Our implementation in GitHub

with standard VAD mapper and 1NN classifier gives about 2 times better accuracy and 4 times better macro F1 compared to the corresponding model with MSE loss. This gap may indicate that MAE is a more appropriate metric for the VAD space than MSE.

Moreover, The model based on MAE+CE loss gave similar results to the other models. It achieved slightly better accuracy performance compared to the simpler model that used MAE loss, but yielded lower Macro F1 performance.

4.4 Experimenting Arbitrary Mapper

To test whether VAD really gives our model insights, we present an experiment in which we use a *VAD mapper* (3.1) that does not rely on VAD. This mapper maps arbitrarily the 28 emotions to $[0, 1]^3$ space.
We want to use a mapping that spreads our 28 emotions as evenly as possible in space, with the view that a uniform distribution will lead to better learning of the model. The mapper we chose is based on *Arrangement of points on sphere* (Pfoertner, 2005). We use this 28 point arrangement as the points in the $[0, 1]^3$ space corresponding to the 28 emotions as our mapper. It is important to point out that the mapping between the emotions and the points is arbitrary, in order to examine whether VAD, our non-arbitrary choice, did indeed contribute to the model.

VAD mapper	Accuracy	Macro F1
NRC VAD Scaled (Our model)	0.59	0.454
Arbitrary-Uniform (Pfoertner)	0.591	0.454

Table 3: Comparison of the classification results of a model using scaled NRC VAD mapper with a model using arbitrary mapping. In both models we use MAE loss.

Surprisingly, as we can see from Table 3, the arbitrary-uniform mapper gave similar performance compared to an analogous model that used VAD-based mapper. We believe that our model did not utilize the VAD space as intended, but saw each point in the VAD space similarly to a simple emotion label. This may also be one of the reasons why our model achieved results comparable to the results obtained in the classical classification model, as we seen in section 4.3.

5 Transfer Learning Experiments

In the following section we experiment our model's regressor component[12] on multiple datasets, utilizing the predicted VAD (as a black-box) to predict sentiments and emotions on new datasets. The experiments' prediction settings are either zero-shot or based on simple statistical models, in order to empirically challenge our VAD regression.

[12]We use the model *VAD Scaled, MAE, SVM*

5.1 IMDB Reviews Dataset Experiment

- **Data:** *IMDB Reviews* is a dataset (Maas et al., 2011) with (positive / negative) sentiment for each movie review.
- **Experiment:** Our experiment on this dataset is done in a **zero-shot** setting. We use the VAD predicted by our model, and specifically extract the 'Valence' dimension, in order to determine the sentiment of the given review (with some threshold). We base this prediction method on the fact the valence stands for the positiveness of the text.
- **Results:** 69.71% accuracy on the test set.

5.2 Emotion Dataset Experiment

- **Data:** *Emotion* is a dataset (Saravia et al., 2018) with emotion labels (from: joy, love, anger, fear).
- **Experiment:** We use the predicted VAD to train a **simple SVM** classifier (that classifies VAD to one of these emotions), we then use this classifier to predict the emotions of the test set predicted VADs.
- **Results:** 57% accuracy on the test set.

We can deduce that the regressor indeed correctly extract the VAD in most cases, as it outperform the "random baseline" in both cases (50% and 25% accordingly). However, the results are not remarkable and can possibly be achieved using simple BOW (Bag Of Words) models.

6 Conclusion

In this paper we presented a new approach for fine-grained sentiment analysis of text, based on VAD representation. Our model consists of a regression model that evaluates a point in the VAD space that represents the emotion expressed in a given sentence, and a classifier that converts this VAD point to a discrete emotion.

We presented experiments that test the quality of our regressor and the complete model, in which we examined different variants of our model.

From the experiments we obtained that our model achieves results comparable to the baseline model, but is not able to further exceed those baseline results.

It could be argued that one of the reasons that our model yields similar results to the classic classification model is that it doesn't fully utilize the emotional knowledge embodies by VAD, as we would expect it. An initial evidence for this may be the arbitrary mapper experiment (4.4) in which we saw that using authentic VAD as the base for our mapper does not benefit the model compared to an arbitrary mapping.

Another explanation for these results might be that a sophisticated transformer-based model such as BERT is strong enough to learn an effective representation of emotions by itself, so the additional knowledge we allegedly gave to it did not contribute. In other words,

the fact that inserting BERT with psychological expertise does not improve results drastically is an evidence for BERT's strengh.

Finally, while the fact that our models do not dramatically outperform the baseline can be seen as a disappointment for VAD representation, we believe that they leave an open door to other regression methods that might utilize a similar approach. Moreover, as we showed, such regression methods can also be used for new tasks achieving non-trivial results.

For future work, we suggest further exploring different / higher dimensional (i.e. with dimension bigger than 3) representation of emotions, possibly by using the framework we proposed in this paper. In addition, we believe that this kind of work should be done in valuable collaboration with psychological researchers.

References

[Alvarez-Gonzalez et al.2021] Nurudin Alvarez-Gonzalez, Andreas Kaltenbrunner, and Vicenç Gómez. 2021. Uncovering the limits of text-based emotion detection. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2560–2583, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

[Cowen et al.2019] Alan Cowen, Disa Sauter, Jessica L. Tracy, and Dacher Keltner. 2019. Mapping the passions: Toward a high-dimensional taxonomy of emotional experience and expression. *Psychological Science in the Public Interest*, 20(1):69–90. PMID: 31313637.

[Demszky et al.2020] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online, July. Association for Computational Linguistics.

[Ekman1992] Paul Ekman. 1992. An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200.

[Maas et al.2011] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June. Association for Computational Linguistics.

[Mohammad2018] Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia, July. Association for Computational Linguistics.

[Park et al.2021] Sungjoon Park, Jiseon Kim, Seonghyeon Ye, Jaeyeol Jeon, Hee Young Park, and Alice Oh. 2021. Dimensional emotion detection from categorical emotion. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4367–4380, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

[Pfoertner2005] Hugo Pfoertner. 2005. Arrangement of points on sphere, available at http://www.randomwalk.de/sequences/a084824.txt. http://www.pfoertner.org/.

[Russell and Mehrabian1977] James Russell and Albert Mehrabian. 1977. Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11:273–294, 09.

[Saravia et al.2018] Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium, October-November. Association for Computational Linguistics.

[Singh et al.2021] Gargi Singh, Dhanajit Brahma, Piyush Rai, and Ashutosh Modi. 2021. Fine-grained emotion prediction by modeling emotion definitions. *CoRR*, abs/2107.12135.

7 Appendix

7.1 MAE + CE Loss Formal Definition

For a given VAD predictions $\hat{Y} : \hat{Y_1}, \hat{Y_2}, ..., \hat{Y_n}$, with VAD targets $Y : Y_1, Y_2, ..., Y_n$, and one-hot label targets[13] $L : L_1, L_2, ..., L_n$, The **MAE + CE Loss** is defined as follows:

$$MAE(\hat{Y}, Y) + \lambda \cdot CE(softmax(\phi(\hat{Y})), L) \quad (1)$$

where:

$$\phi(\hat{Y}) := -\|\hat{Y_i} - E_j\|_1$$

where $E_1, E_2, ..., E_{28}$ are the VAD points of our 28 emotions, and λ was chosen by tuning[14].

Intuitively, emotions that are closer in the VAD space to the point that the model predicted, will have a higher probability in the distribution (over the emotions) we get from $softmax(\phi(\hat{Y}))$.

[13] $(L_i)_j$ is equal to 1 if j is the emotion label of input i and 0 otherwise.

[14] We chose $\lambda = 0.1$.