LDK 2023

Unifying Emotion Analysis Datasets using Valence Arousal Dominance (VAD)

Ryutaro Takanami
Lancaster University, UK
ryu.takanami212@gmail.com

Mo El-Haj
Lancaster University, UK
m.el-haj@lancaster.ac.uk

Abstract

This paper presents a novel approach to unifying various emotional datasets in Natural Language Processing (NLP) using the Valence Arousal Dominance (VAD) framework. Emotion analysis, which aims to deeply analyse emotions and understand user behaviour, is a complex research area that requires large, standard, and unified datasets. However, the lack of such datasets in NLP has been a challenge in advancing the field. Our approach maps diverse emotions from different datasets into four categories: joy, anger, fear, and sadness using the VAD framework. This process creates multidimensional emotional scores that are consistent across datasets, regardless of the number of emotions included. By unifying these datasets, we were able to train a BERT model on the combined data and improve the performance of emotion detection.

1 Introduction

Emotion detection is a crucial aspect of Natural Language Processing (NLP). There are two main approaches used in NLP for emotion detection: the categorical model and the dimensional model. The categorical model, based on the work of Ekman and Plutchik (Ekman, 1999; Plutchik, 1980), suggests that human emotions can be represented as basic emotions such as joy, sadness, and anger. On the other hand, the dimensional model, based on the work of Russell et al. (Russel, 1980), proposes that emotions can be captured as a point in a multidimensional space, with unconscious elements driving categorical feelings.

While the categorical model provides a straightforward approach to capturing emotions, it has some limitations. For example, it assumes that emotions are discrete categories, and fails to account for the possibility of ambiguity or mixed emotions. The dimensional model overcomes these limitations by representing emotions as points in a multidimensional space, allowing for the possibility of mixed or ambiguous emotions.

Despite the advantages of the dimensional model, there are still challenges in emotion detection. One of the significant obstacles is the lack of standardised emotional datasets. The available datasets differ in terms of the number of emotions and the types of emotions annotated, making it challenging to train a single machine learning model. To tackle this issue, we propose a method of unifying annotations from different datasets using Valence Arousal Dominance (VAD) to convert labels into a unified VAD score that represents emotions in a 3-dimensional space. This approach provides a more comprehensive understanding of emotions and maximises the use of available datasets to train machine learning models.

In addition to unifying annotations, we address the issue of "weak emotions" by annotating such instances with a neutral VAD score. Sentences that contain conflicting emotions or those that do not exhibit a clear or strong emotional response are referred to as weak emotion sentences. Conventional annotation methods treat sentences with the same emotion equally, but VAD can detect and provide a more nuanced label by assigning a score range instead of a fixed annotation value.

This study has three main objectives:

1. To provide a flexible mapping model that can incorporate different types of emotions from different datasets and unify them into a polarity score of four emotions: joy, anger, fear, and sadness.

2. To improve the accuracy of emotion prediction compared to sentiment polarity detection.

3. To investigate whether the VAD scores can detect neutrality, or what we later refer to as 'weak emotions'.

In conclusion, our approach to emotion detection provides a more nuanced understanding of emotions in text and helps to overcome some of the limitations of existing methods. By unifying annotations using VAD, we can train machine learning models with greater accuracy and provide more comprehensive insights into the emotions expressed in text.

2 Related Work

One of the earliest emotion detection approaches was the use of lexicons, pre-defined dictionaries of words and their associated emotional valence (Mohammad, 2018). This approach is simple and straightforward, but it is limited by the size and scope of the lexicon, as well as by the fact that words can have multiple meanings and connotations.

Another approach to emotion detection is the use of machine learning algorithms, which can learn to identify patterns in data and predict emotions expressed in text (Pang and Lee, 2004; El-Haj et al., 2016). However, machine learning algorithms require large amounts of labeled data to train effectively, and the lack of standardised emotion datasets has hindered progress in this field. To address this challenge, researchers have proposed unifying different emotion datasets to create a larger, more comprehensive dataset for training machine learning models (Mohammad, 2018; Abdul-Mageed and Ungar, 2017). By mapping varied emotions from different datasets into a common set of categories, these unified datasets can provide a more nuanced understanding of emotions in text, while also allowing for more accurate predictions of emotions.

Other approaches have been proposed to improve emotion detection in text, such as the use of lexicons, pre-defined dictionaries of words and their associated emotional valence (Mohammad, 2018). Another approach is the use of machine learning algorithms, which can learn to identify patterns in data and predict emotions expressed in text (Pang and Lee, 2004). However, machine learning algorithms require large amounts of labeled data to train effectively, and the lack of standardised emotion datasets has hindered progress in this field (Alwakid et al., 2022).

In recent years, there has been a growing interest in using the Valence Arousal Dominance (VAD) model as a way to detect and unify different emotion datasets (Kulkarni and Bhattacharyya, 2021; Luengo et al., 2010). The VAD model captures the affective quality of emotions and offers a more nuanced understanding of emotions in different contexts (Russel, 2003). By mapping different emotions to a common set of VAD scores, researchers can create a unified dataset that is more comprehensive and offers a more nuanced understanding of emotions in text. This approach has the potential to improve the accuracy of emotion detection algorithms and provide a more fine-tuned understanding of emotions expressed in text. To address this challenge, we propose unifying different emotion datasets using VAD, a multidimensional model of emotions that captures valence, arousal, and dominance. By mapping varied emotions from different datasets into four categories - joy, anger, fear, and sadness - we can create multidimensional emotional scores that work across different datasets, regardless of the number of emotions introduced in each. This approach enables us to train machine learning models on a unified dataset, which can improve emotion detection performance and provide more comprehensive insights into the emotions expressed in text.

3 Datasets

This research uses five different datasets mainly focusing on text written in English. Four of the studied datasets are annotated with coarse-grained categorical emotions, while the fifth has VAD labels.

3.1 Stance Sentiment Emotion Corpus (SSEC)

The Stance Sentiment Emotion Corpus (SSEC) is an annotation of the SemEval-2016 Task 4[1] Twitter stance. The corpus contains 4,870 tweets, each paired with eight emotional categories: Anger, Anticipation, Disgust, Fear, Joy, Sadness, Surprise, and Trust. Each tweet was annotated by three to six annotators who were undergraduate students of media computer science (Schuff et al., 2017). SSEC is a widely used dataset in the emotion detection field, and its focus on stance and emotions in tweets makes it particularly relevant to social media analysis.

3.2 SemEval-2018 Task 1 EC

SemEval-2018 Task 1 EC is a dataset of 3,259 English tweets paired with 11 categorized emotion

[1] SemEval-2016 Task 4: Sentiment Analysis in Twitter: http://alt.qcri.org/semeval2016/task4/

labels: Anger, Anticipation, Disgust, Fear, Joy, Love, Optimism, Pessimism, Sadness, Surprise, and Trust (Mohammad et al., 2018). The dataset was created by having seven annotators label one or more emotions that represent the tweeter's emotion from a sentence. This dataset is especially valuable for research that focuses on microblogging sites such as Twitter.

3.3 WASSA-2017 Shared Task on Emotion Intensity (WASSA)

WASSA-2017 is a dataset containing about 4,636 manually annotated tweets, categorized into four emotions: Anger, Fear, Joy, and Sadness (Mohammad and Bravo-Marquez, 2017). The authors gathered tweets containing emotional words representing each category. The emotional words were chosen using Roget's Thesaurus (Chapman et al., 1977). The tweets were manually annotated using crowd-sourcing. WASSA-2017 is a useful dataset for emotion detection research because of its focus on emotion intensity.

3.4 SemEval-2017 Task 4 A (Polarity)

SemEval-2017 Task 4 A is a dataset from the Sentiment Analysis in Twitter challenge (Rosenthal et al., 2017). It contains 11,906 polarity-emotion annotated tweets, with polarity labels of "positive," "neutral," and "negative." Tweets that mentioned any internationally trending events on Twitter were chosen for data collection, and the tweets were annotated with 3-point scales (positive, neutral, and negative) (Rosenthal et al., 2017). This dataset is valuable for research that focuses on sentiment analysis and emotion detection.

3.5 EmoBank

EmoBank is a dataset containing 10,062 sentences paired with continuous VAD labels (Buechel and Hahn, 2017). It is the largest VAD-model text corpus to the best of our knowledge. The sentences were extracted from several online sources, such as blogs, essays, news headlines, and tweets. The dataset was annotated with 5-point scales (ranging from 1 to 5) by crowd workers (Buechel and Hahn, 2017). EmoBank is a valuable resource for emotion detection research because of its large size and its fine-grained VAD labels.

4 Pre-processing

In this section, we detail the pre-processing steps for the training set that will be used as input for our BERT model.

The BERT model is trained to predict VAD values and to convert these values into categorical labels, based on the required emotion categories. For datasets, such as SemEval-2018 and SSEC (Section 3.1), which are annotated with multiple categorical emotions in a single sentence, we average the VAD values of each emotion to obtain the overall VAD value of that sentence before BERT model training. This is because the VAD value of a sentence should consist of only one score for the training of the machine learning BERT model. For instance, if a sentence is labeled with "joy", "love", and "trust", the VAD scores for each will be something like: joy" = [980, 824, 794], "love" = [1000, 519, 673] and "trust" = [888, 547, 741]. The score of the sentence will then become a three-dimensional score of: Valence V = (980+1000+888)/3 = 956, Arousal A = (824+519+547)/3 = 630, and Dominance D = (794+673+741)/3 = 736.

In the SemEval-2018 and SSEC datasets, multiple labels can be assigned to a single sentence, but not if it is considered neutral. To account for this, we set the intermediate value in VAD space, 500, for sentences without any labels. This is because the range of each axis is a VAD score from 0 to 1000, and in this research, we choose 500 as the moderate strength of the emotion score, or what can be considered as no emotion but falls within the neutral score range, as we demonstrate later in Experiment 2 (Section 5.2).

For the EmoBank dataset (Section 3.5), the pre-existing VAD values range between 1 and 5 points, which is different from our VAD scale. In this work, we use a scale of 0 to 1000 for our VAD score annotations, as the NRC VAD lexicon (Mohammad, 2018) adopted the same scale. To transform the categorical labels in EmoBank to our scale of 0-1000 VAD scores, we use the following formula, where EmoBank-Score is the 1-5 Likert scale score given by the human annotators:

$$VADScore = (EmoBank-Score-1)/4*1000 \quad (1)$$

We also pre-process the text of the datasets. The majority of the sentences in the datasets are sourced from Twitter, so we pre-process the data by removing mentions and URLs, as they are considered unrelated to expressing emotions. On the other hand, hashtags are retained, as they can help capture cases where emotions are directly included in

the hashtag, such as "#love".

5 Experimental Work

The experimental work is divided into two phases. In the first phase, we train a BERT machine learning model to predict categorical emotions from the unified representation of multiple datasets using the VAD model. In the second phase, we demonstrate how the model can be adapted to capture what we refer to as "weak emotions" which are neutral emotions found in sentiment datasets such as SemEval-2017 (Section 3.4).

5.1 Experiment 1: Predicting Categorical Emotions

This experiment addresses the first two objectives of the research as outlined in the Introduction (Section 1).

In this experiment, we create a combined prediction model from multiple differently annotated datasets and evaluate if the accuracy can be improved compared to training on individual datasets. The combined model was trained on the EmoBank, SemEval-2018, and SSEC datasets (denoted as "All"). Additionally, separate models were trained for each individual dataset (denoted as "Emo", "Sem", and "SSEC", respectively), as shown in Table 1[2].

We use the WASSA dataset (Section 3.3) as the test set for this experiment, as each sentence in WASSA is annotated with a single categorical label (joy, anger, fear, or sadness), making it an appropriate dataset to evaluate our models. The results of the BERT model are expressed in terms of VAD scores and are labeled according to the WASSA categories for comparison. This is done by calculating the Euclidean Distance between the predicted VAD scores and the VAD scores of each of the four emotions as labeled in WASSA, and the emotion with the minimum distance becomes the predicted label for a given sentence.

5.2 Experiment 2: Detecting Weak Emotions

This experiment addresses the third objective of the research by investigating whether the VAD scores can detect neutrality (weak emotions).

For this experiment, we use the SemEval-2017 dataset as the testing set, as it has a polarity annotation of positive, neutral, and negative emotions. The Valence dimension ("V" axis) in VAD is used to predict the polarity emotions. Valence is known to be the most stable dimension in VAD space, where individual perceptions are represented (Hoffman et al., 2012).

We use the VAD score prediction models trained in Experiment 1 to predict the polarity emotions by using SemEval-2017 as the test data. Before comparing the results to the true labels, the predictions are visualised in a scatter plot to show how the combination of multiple datasets increases the representation of emotions estimated by the BERT model (Section 6.2). After predicting the sentiment of a sentence in dimensional space, we convert the predicted V score into categorical emotion labels: positive, neutral, and negative.

Since the test data is annotated with categorical variables, we need to change the predicted V-values, represented by the V-dimension, to categorical values. To do this, we set polarity emotion thresholds for the V-dimension at 300 and 700. It seems reasonable to classify emotions less than 300 as negative, emotions between 300 and 700 as neutral, and emotions above 700 as positive, dividing the V-Score range of 0-1000 into three semi-equal ranges.

6 Results and Evaluation

6.1 Experiment 1

The results of the emotion prediction accuracy for the four emotions (joy, anger, fear, and sadness) tested using the WASSA dataset are shown in Table 1. The results demonstrate that training the BERT model on a combination of different emotion-based datasets (denoted as 'All') produces results that are equivalent to training using a single dataset. This suggests that mapping the differently annotated datasets is capable of producing comparable results, and the combination of different datasets did not result in a decrease in accuracy. In particular, when some of the models trained individually (denoted as SEEC) had lower accuracy, the combination of several datasets helped the BERT model learn better how to predict emotions.

	All	SEEC	Emo	Sem
Four emotions	0.44	0.25	0.41	0.45

Table 1: Emotion prediction accuracy.

The number of sentences per emotion is shown

[2]WASSA and SemEval-2017 datasets are used as testing sets and were therefore not included in the training process

in Table 2 and Figure 1. The imbalance in the data resulted in a bias in emotion prediction, which is expected since anger and joy are the most frequent classes. This can be seen in the results of the models by emotion, shown in Table 3. As a potential solution, future experiments could reduce the number of emotions and increase emotions that are close in the VAD space (e.g., fear and sadness).

	All	SEEC	Emo	Sem
anger	10555	1997	7734	824
joy	3966	1472	1091	1403
fear	2265	1324	270	671
sadness	1405	77	967	361

Table 2: Number of sentences by emotion.

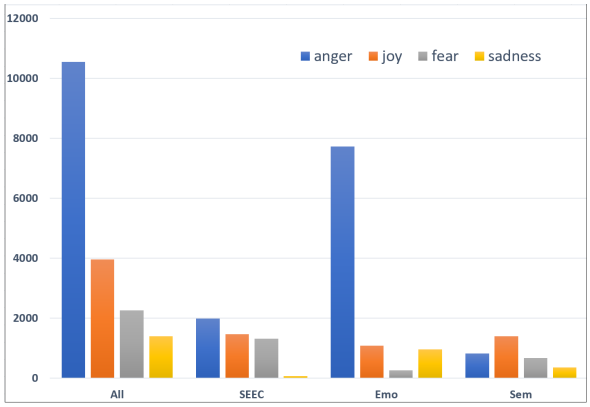


Figure 1: Number of sentences by emotion.

	All	SEEC	Emo	Sem
anger	0.75	0.10	0.96	0.68
joy	0.76	0.34	0.39	0.90
fear	0.12	0.34	0.01	0.15
sadness	0.01	0.34	0.01	0.01

Table 3: Prediction accuracy by emotions.

6.2 Experiment 2

To illustrate that combining datasets has increased the range of emotions that the models can predict, we show a scatter plot of the predictions for each model in Figure 2. The colours in the plot represent the correct label prediction: red for positive, yellow for neutral, and blue for negative. The Y axis is the ID of the predicted sentence, and the X axis is the V score range. None of the models trained on a single dataset were able to categorise all three categories.

It can be seen that the All Model has the richest variety of emotions to predict and is better able to pick up subtle differences in emotions. Moreover, the All Model plot confirms that our threshold values for the V-dimension are reasonable, as the V-score seems to be divided into three categories between around 300 and 700.

	All	SEEC	Emo	Sem
Positive	0.494	0.411	0.0	0.494
Neutral	0.587	0.0	0.482	0.0
Negative	0.571	0.5	0.0	0.442
Average	0.551	0.304	0.161	0.312

Table 4: Accuracy of polarity emotions.

The prediction accuracy of each model for the three categories (positive, negative, and neutral) is examined in Table 4. In terms of prediction accuracy, the All Model has the highest accuracy, demonstrating that the BERT model was able to learn better when a combination of several emotion-based datasets was used. None of the models trained on a single dataset were able to categorise all three categories with consistent accuracy, as confirmed by the scatter plots in Figure 2.

7 Conclusion

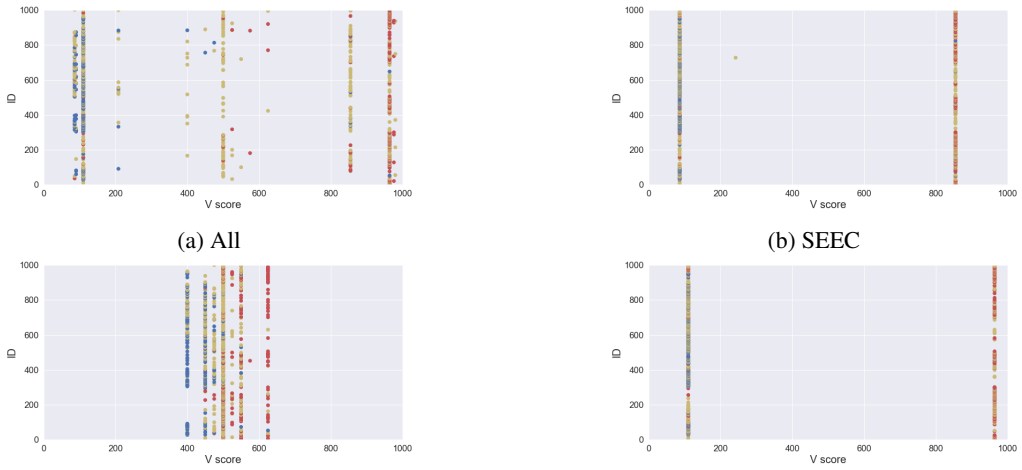
The results of Experiments 1 and 2 in this study demonstrate the benefits of training with larger emotion-based datasets. By transforming these datasets using the Valence Arousal Dominance (VAD) framework, our findings suggest that it is possible to predict a wider range of emotional expressions. The results of the polarity analysis in Experiment 2 further support this conclusion.

As future work, it is expected that increasing the number of datasets used in training will result in improved accuracy of emotion prediction. The experiments conducted in this study also showed that it is possible to predict weak emotions, which are often overlooked by conventional sentiment analysis models.

References

Muhammad Abdul-Mageed and Lyle Ungar. 2017. Emonet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of ACL'17*, pages 718–728.

Ghadah Alwakid, Taha Osman, Mahmoud El Haj, Saad Alanazi, Mamoona Humayun, and



(a) All

(b) SEEC

(c) EMOBANK

(d) SemEval2018

Figure 2: Predicted V scores.

Najm Us Sama. 2022. Muldasa: Multifactor lexical sentiment analysis of social-media content in nonstandard arabic social media. *Applied Sciences*, 12(8):3806.

Seven Buechel and Udo Hahn. 2017. Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis.

Robert. L. Chapman, Peter. Mark. Roget, et al. 1977. *Roget's international thesaurus*. Crowell.

P. Ekman. 1999. *Basic Emotions. In Dalgleish, Tim and Powers, M. J. (eds.), Handbook of Cognition and Emotion*, volume 1. Wiley.

Mahmoud El-Haj, Paul Edward Rayson, Steven Eric Young, Martin Walker, Andrew Moore, Vasiliki Athanasakou, and Thomas Schleicher. 2016. Learning tone and attribution for financial text mining.

Holger Hoffman, Andreas Scheck, Timo Schuster, Steffen Walter, Kerstin Limbrecht, Harald C. Traue, and Henrik Kessler. 2012. Mapping discrete emotions into the dimensional space: An empirical approach. In *IEEE SMC'12*, pages 3316–3320.

Manasi Kulkarni and Pushpak Bhattacharyya. 2021. Retrofitting of pre-trained emotion words with vad-dimensions and the plutchik emotions. In *Proceedings of ICON'21*, pages 529–536.

Iker Luengo, Eva Navas, Igor Odriozola, Ibon Saratxaga, Inmaculada Hernaez, Inaki Sainz, and Daniel Erro. 2010. Modified ltse-vad algorithm for applications requiring reduced silence frame misclassification. In *LREC*.

M. Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words.

S. Mohammad and F. Bravo-Marquez. 2017. Wassa-2017 shared task on emotion intensity.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. *arXiv preprint cs/0409058*.

R. Plutchik. 1980. *Emotion: Theory, research, and experience*, volume Vol. 1, Theories of emotion. Academic Press.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. Semeval-2017 task 4: Sentiment analysis in twitter.

J. Russel. 1980. A circumplex model of affect.

J. Russel. 2003. Core affect and the psychological construction of emotion.

Hendrik Schuff, Jeremy Barnes, Julian Mohme, Sebastian Padó, and Roman Klinger. 2017. Annotation, modelling and analysis of fine-grained emotions on a stance and sentiment detection corpus.