ABSTRACT

We propose VADEC, a multi-task framework that exploits the correlation between the categorical and dimensional models of emotion representation for better subjectivity analysis. Focusing primarily on the effective detection of emotions from tweets, we jointly train multi-label emotion classification and multi-dimensional emotion regression, thereby utilizing the inter-relatedness between the tasks. Co-training especially helps in improving the performance of the classification task as we outperform the strongest baselines with 3.4%, 11%, and 3.9% gains in Jaccard Accuracy, Macro-F1, and Micro-F1, respectively. We further justify the utility of our proposed multi-task approach.

CSC CONCEPTS

- Information systems → Sentiment analysis.

KEYWORDS

Coarse-grained Emotion Analysis; Fine-grained Emotion Analysis; Valence-Arousal-Dominance; Multi-task Learning; Twitter; COVID

1 INTRODUCTION

With the proliferation of social media, as more and more people express their opinions online, detecting human emotions from their written narratives, especially tweets has become a crucial task given its widespread applications in e-commerce, public health monitoring, disaster management, etc. [17, 18]. Categorical models of emotion representation such as Plutchik’s Wheel of Emotion [21] or Ekman’s Basic Emotions [8] classify affective states into discrete categories (joy, anger, etc.). Dimensional models on the other hand describe emotions relative to their fundamental dimensions. Russel and Mehrabian’s VAD model [23] for instance interprets emotions as points in a 3-D space with Valence (degree of pleasure or displeasure), Arousal (degree of calmness or excitement), and Dominance (degree of authority or submission) being the three orthogonal dimensions. Accordingly, the literature on text-based emotion analysis can be broadly divided into coarse-grained classification systems [10, 12–14, 28] and fine-grained regression systems [22, 24, 29, 30]. Although a coarse-grained approach is better-suited for the task of detecting emotions from tweets as observed in [4], prior works fail to exploit the direct correlation between the two models of emotion representation for finer interpretation. We utilize the better representational power of dimensional models [4] to improve the emotion classification performance by proposing VADEC that jointly trains multi-label emotion classification and multi-dimensional emotion regression in a multi-task framework.

Multi-task learning [6] has been successfully used across a wide spectrum of NLP tasks including emotion analysis [1, 30]. While AAN [30] takes an adversarial approach to learn discriminative features between two emotion dimensions at a time, All_In_One [1] proposes a multi-task ensemble framework to learn different configurations of tasks related to coarse- and fine-grained sentiment and emotion analysis. However, none of the methods combine the supervisions from VAD and categorical labels. Our proposed framework (Section 2) consists of a classifier module that is trained for the task of multi-label emotion classification, and a regressor module that co-trains the regression tasks corresponding to the V, A, and D dimensions. Owing to the unavailability of a common annotated corpus, the two tasks are trained using supervisions from their respective benchmark datasets (reported in Section 3.1), which further justifies the utility of our proposed multi-task approach. VADEC learns better shared representations by jointly training the two modules, that especially help in improving the performance of the classification task, thereby achieving state-of-the-art results on the AIT [17] and SenWave [27] datasets (Section 3.3). For the
We discover the major factors contributing towards the various emotions and find their trends to correlate with real-life events. Since we primarily focus on detecting emotions from tweets, we use Multi-label Emotion Classifier and Multi-dimensional Emotion Regressor when trained individually. (c) VADEC represents our Multi-Task Affect Classifier that co-trains the two modules by optimizing the joint loss.

### 2 VADEC ARCHITECTURE

Figure 1 illustrates the architecture of VADEC, that jointly trains a multi-label emotion classifier and a multi-dimensional emotion regressor with supervision from their respective datasets. Since we primarily focus on detecting emotions from tweets, we use BERTweet [19] to serve as our text-encoder. It is shared by the two modules and is hereby referred to as the shared layer. The 768-dim. [CLS] token embedding of the sentence/tweet obtained from BERTweet is first passed through a fully connected (FC) layer with 256 neurons in both the modules respectively. The classifier passes this intermediate representation through another FC layer with 11 output neurons, each activated using Sigmoid with a threshold of 0.5 to predict the presence/absence of one of the 11 emotion categories. Binary Cross-Entropy (BCE) with L2-norm regularization is used as the loss function, hereby referred to as the EC_Loss. Similarly, the regressor passes the 256-dim. intermediate representation through an FC layer with 3 output neurons (with Sigmoid activation) corresponding to the V, A and D dimensions. It then jointly optimizes the Mean Squared Error (MSE) loss of all three dimensions, hereby referred to as the VADR_Loss. VADEC jointly trains the two modules by optimizing the following multi-task objective:

\[
VADEC_{Loss} = \lambda \cdot EC_{Loss} + (1 - \lambda) \cdot VADR_{Loss}
\]

Here, \(\lambda\) represents a balancing parameter between the two losses. The weighted joint loss backpropagates through the shared layer, thereby fine-tuning the BERTweet parameters end-to-end.

### 3 RESULTS AND DISCUSSION

#### 3.1 Datasets

For our experiments, we consider EMOBANK, a VAD dataset, and two categorical datasets, AIT and SenWave as described below:

- **EMOBANK** (Buechel and Hahn [5]) : A collection of around 10k English sentences from multiple genres (8,062 for training, and 1K sentences each for validation and testing), each annotated with continuous scores (in the range of 1 to 5) for Valence, Arousal, and Dominance dimensions of the text.

- **AIT** (Mohammad et al. [17]) : Created as part of SemEval 2018 Task 1: “Affect in Tweets”, it consists of 10,983 English tweets, 6,838 for training, 886 for validation, 3,259 for testing, each with labels denoting the presence/absence of a total of 11 emotions.

- **SenWave** (Yang et al. [27]) : Till date the largest fine-grained annotated COVID-19 tweets dataset consisting of 10K English tweets (8K for training, and 1K each for validation and testing), each with corresponding labels denoting the presence/absence of 11 different emotions specific to COVID-19.

#### 3.2 Experimental Setup

For all our model variants, we perform extensive experiments with different sets of hyper-parameters and select the best set w.r.t. lowest validation loss. Before evaluating the performance on the test set, we combine the training and validation data and re-train the models with the best obtained set of hyper-parameters (learning rate = 2e-5, weight decay = 0.01, \(\lambda = 0.5\), and no. of epochs = 5 for VADEC). For the regression task, the outputs of Sigmoid activation at each of the three output neurons are suitably scaled before calculating the MSE loss since the ground-truth VAD scores are in the range of 1-5. As model ablations, we investigate the role played by features derived from affect lexicons by additionally appending a 194-dim. Empath\(^1\) feature vector to the intermediate representations learnt by our model variants to be used for final predictions. Parameters of our shared encoder are initialized with pre-trained model weights (roberta-base for RoBERTa, and bertweet-base for BERTweet) from the HuggingFace Transformers library [25]. Other model parameters are randomly initialized. All our model variants are trained end-to-end with Adam\(^2\) optimizer [16] on Tesla P100-PCIE (16GB) GPU. We additionally ensure the reproducibility of our results and make our code repository\(^2\) publicly accessible.

#### 3.3 Evaluating Emotion Classification

We first discuss the comparative results of our model variants and ablations on the AIT dataset. We then respectively report our state-of-the-art results achieved on the AIT and the SenWave datasets.

---

1. https://github.com/Ejhfast/empath-client
2. https://github.com/atharva-naik/VADEC
AIT Dataset

As metrics we use Jaccard Accuracy, Macro-F1, and Micro-F1 [17]. Among recent baselines: (i) BERTL (Park et al. [20]) denotes the scores obtained by fine-tuning BERT-Large [7] on the AIT dataset, and (ii) NTUA-SLP (Baziotis et al. [3]) represents the winning entry for this (sub)task of SemEval 2018 Task 1 [17], where the authors take a transfer learning approach by first pre-training their Bi-LSTM architecture, equipped with multi-layer self attentions, on a large collection of general tweets and the dataset of SemEval 2017 Task 4A, before fine-tuning their model on this dataset. Among our model variants and ablations: (i) EC represents our classifier module, when trained as a single task (Fig. 1a), (ii) EC_{RoBERTa} uses RoBERTa [15] instead of BERTweet as the shared layer.

From Table 1, NTUA-SLP surprisingly outperforms BERTL (on Jac. Acc. and Micro-F1), a heavier model with 336M parameters. EC (trained with BERTweet) comfortably beats EC_{RoBERTa} demonstrating the better efficacy of BERTweet in learning features from tweets. The sparse Empath feature vectors do not however add any value to the rich 768-dim. contextual representations learnt using BERT-based methods. We obtain our best results with VADEC, with respectively 3.4% and 3.9% gains in Jac. Acc., and Micro-F1 over NTUA-SLP, and 11% gain in Macro-F1 over BERTL.

SenWave Dataset

Considering the superior performance of VADEC over all its model variants and ablations from Table 1, here we directly compare the results of VADEC, re-trained with SenWave [27], with the ones reported by the authors of [27], serving as the only available baseline on this dataset. Following [27], we use Label Ranking Average Precision (LRAP), Hamming Loss, and Weak Accuracy (Accuracy) as metrics in addition to the ones reported in Table 1. As observed from Table 2, VADEC achieves SOTA by outperforming the baseline scores with 11.3% performance gain averaged over all 6 metrics.

Overall, our results from Tables 1 and 2 demonstrate the advantage of utilizing the VAD supervisions for improving the performance of the multi-label emotion classification task.

3.4 Evaluating Emotion Regression

Pearson Correlation Coefficient $r$ is used as the evaluation metric for this task. All the models are evaluated on the EMOBANK corpus. Among recent baselines: (i) AAN (Zhu et al. [30]) employs adversarial learning between two attention layers to learn discriminative word weight parameters for scoring two emotion dimensions at a time. The authors report the VAD scores for all 6 domains and 2 perspectives of EMOBANK. For comparison, we use their highest correlation score for each dimension, (ii) All In One (Akhtar et al. [1]) represents a multi-task ensemble framework which the authors use for learning four different configurations of multiple tasks related to emotion and sentiment analysis, (iii) SVR-SLSTM (Wu et al. [26]) represents a semi-supervised approach using variational autoencoders to predict the VAD scores, and (iv) BERTL (EB ← AIT) [20], the current state-of-the-art, fine-tunes BERT-Large [7] on the AIT dataset to predict VAD scores by means of minimizing EMD distances between the predicted VAD distributions and sorted categorical emotion distributions as a proxy for target VAD distributions. For comparison, we use their reported scores obtained upon further fine-tuning their best-trained model on the EMOBANK corpus. Our model variants include (i) VADR which represents our regressor module, when trained as a single task (Fig. 1b), (ii) VAD_{RoBERTa} an ablation where we experiment with RoBERTa as the shared layer, (iii) VADEC (AIT), and (iv) VADEC (SenWave) representing the scores of our multi-task model when trained respectively with the AIT and SenWave datasets.

From Table 3, VAD_{RoBERTa} shows the highest correlation (0.511) on the D dimension. VADR (w/ BERTweet) however outperforms VAD_{RoBERTa} on the other two dimensions. Contrary to our observations in the classification task, co-training does not help in improving the performance of the regression task, as can be confirmed from the results of VADEC (AIT) and VADR. Although we are outclassed by BERTL (EB ← AIT) on the A dimension, VADEC (AIT) comfortably outperforms BERTL (EB ← AIT) on the V and D dimensions. VADEC (SenWave) further outclasses both VADEC (AIT) and BERTL (EB ← AIT) on V and D with 7.6% and 16.5% gains respectively. To conclude, although joint-learning does not help the regression task as much as it helps in improving the classification performance (which in fact is our main objective), we still achieve noticeable improvements in majority of emotion dimensions.

3.5 COVID-19 and Indians: A Case Study

For this analysis, we consider Twitter_IN, a subset of COVID-19 Twitter chatter dataset (version 17) [2], containing around 140K English tweets from India posted between January 25th and July 4th 2020. Owing to very few reported cases in India before March 2020, we begin our analysis by predicting emotions from tweets, posted on or after Match 1st 2020, using VADEC trained on EMOBANK dataset.
### Table 4: Few Examples of Single and Multi-label Predictions on Tweets from Twitter_IN

| Tweet                                                                 | Predicted Labels          |
|----------------------------------------------------------------------|---------------------------|
| Let us spare a moment and thought for the junior resident doctors of Mumbai on the frontline fighting it out alone with little help from the government against all odds and at great personal risk | Thankful                  |
| This is the time to fight Covid19 at present but some intelligent Generals are focusing on war and terrorism | Annoyed                   |

#### Multiple Labels

| Tweet                                                                 | Predicted Labels          |
|----------------------------------------------------------------------|---------------------------|
| The first Covid 19 positive from Meghalaya Dr John Sailo Rintathiang passed away early this morning. Sailo 69 who was also the owner of Bethany hospital was tested positive on April 13 2020 | Sad, Official Report      |
| Media is so obsessed with a particular community that they even misspell coronavirus | Annoyed, Joking, Surprise |

### Table 5: Major aspects affecting various emotions among Indians towards the COVID-19 pandemic.

| Emotion  | Major aspects                                                                 |
|----------|-------------------------------------------------------------------------------|
| Annoyed  | govt, politics, death, news, religion, jamaat, work, China, assault, border    |
| Sad      | lockdown, death, distancing, life, family, economy, village, doctor, worker, school |
| Thankful | doctor, service, staff, nurse, app, fund, assistance, leadership             |
| Optimistic| initiative, opportunity, measure, arogyasetuapp, IndiaFightsCorona, stayhome, vaccine, change, support, action |

![Graph](image1.png)  
(a) Annoyed

![Graph](image2.png)  
(b) Optimistic

**Figure 2: Change in Sub-categories of Emotional Triggers towards the COVID-19 pandemic over time.**

and SenWave. Few tweets with their predicted emotions are listed in Table 4. For each emotion, we obtain its contributing aspects by training an unsupervised neural topic model, ABAE (He et al. [11]) on the subset of tweets containing the given emotion as per VADEC predictions. Few emotions along with their most accurate aspects are reported in Table 5. For each emotion, the extracted aspect terms are further filtered and assigned meaningful sub-categories by means of a many-to-many mapping. In Figure 2, we plot the temporal trends of these sub-categories (with roughly equal-sized bins in terms of no. of tweets predicted with the emotion plotted) that respectively made Indians feel annoyed (Fig. 2a) and optimistic (Fig. 2b) over time. In Fig. 2a, the peak in Crowd gathering between March 28th and April 7th can be attributed to the Tablighi Jamaat gatherings3 unfortunately triggering widespread criticism. Fig. 2b shows a high level of Community gratitude in general, with occasional peaks which may be attributed to the events targeted at raising solidarity among the public. For Technology and AI, we observe a peak near the launch date of the Arogya Setu App4 - developed by the Indian Government to identify COVID-19 clusters.

### 4 CONCLUSION AND FUTURE WORK

In this work, we for the first time exploit the correlation between categorical and dimensional models of emotion analysis by proposing VADEC, a multi-task affect classifier with the primary objective of efficiently detecting emotions from tweets. Co-training the tasks of multi-label emotion classification and multi-dimensional emotion regression helps the former thereby achieving state-of-the-art results on two benchmark datasets, AIT (non-COVID) and SenWave (COVID-related). For the regression task, VADEC still outperforms the strongest baseline on the EMOBANK dataset on the V and D dimensions. In future, we would like to investigate the hierarchical relationship between the tasks and analyze the relative impact of each emotion dimension on the emotion classification task.

### ACKNOWLEDGMENTS

This research is partially supported by IMPRINT-2, a national initiative of the Ministry of Human Resource Development (MHRD), India. Niloy Ganguly was partially funded by the Federal Ministry of Education and Research (BMBF), Germany (grant no. 01DD20003).

---

3. https://en.wikipedia.org/wiki/2020_Tablighi_Jamaat_COVID-19_hotspot_in_Delhi
4. https://en.wikipedia.org/wiki/Arogya_Setu
REFERENCES

[1] S. Akhtar, D. Ghosal, A. Ekbal, P. Bhattacharyya, and S. Kurohashi. 2019. All-in-One: Emotion, Sentiment and Intensity Prediction using a Multi-task Ensemble Framework. IEEE Transactions on Affective Computing, 1–1. https://doi.org/10.1109/TAFFC.2019.2926724

[2] Juan M. Banda, Ramya Tekumalla, Guanyu Wang, Jingyuan Yu, Tuo Liu, Yuning Ding, and Gerardos Chowell. 2020. A large-scale COVID-19 Twitter chatter dataset for open scientific research - an international collaboration. ArXiv (2020). https://doi.org/10.5281/zenodo.3900903

[3] Christos Bziotis, Athanasia Nikolaos, Alexandra Chronopoulou, Athanasia Kolovou, Georgios Paraskevopoulos, Nikolaos Ellinas, Shinkanth Narayanam, and Alexandros Potamianos. 2018. NTUA-SLP at SemEval-2018 Task 1: Predicting Affective Content in Tweets with Deep Attentive BNNs and Transfer Learning. In Proceedings of the 12th International Workshop on Semantic Evaluation. Association for Computational Linguistics, New Orleans, Louisiana, 245–255. https://doi.org/10.18653/v1/S18-1037

[4] Sven Buechel and Udo Hahn. 2016. Emotion Analysis as a Regression Problem – Dimensional Models and Their Implications on Emotion Representation and Metrical Evaluation. In Proceedings of the Twenty-Second European Conference on Artificial Intelligence (The Hague, The Netherlands) (EC AI’16). IOS Press, NLN, 1114–1122. https://doi.org/10.3233/978-1-61499-672-9-1114

[5] Sven Buechel and Udo Hahn. 2017. EmoBank: Studying the Impact of Annotation Perspective and Representation Format on Dimensional Emotion Analysis. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics Volume 2: Short Papers. Association for Computational Linguistics, Valencia, Spain, 578–585. https://www.aclweb.org/anthology/ E17-2092

[6] R. Caruana. 2004. Multitask Learning. Machine Learning 28 (2004), 41–75.

[7] 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). Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[8] P. Ekman. 1992. An argument for basic emotions. Cognition & Emotion 6 (1992), 169–200. https://doi.org/10.1080/02699939208410408

[9] Ethan Fast, Binbin Chen, and Michael S. Bernstein. 2016. Empath: Understanding Emotions and Their Implications on Emotion Representation and Metrical Evaluation. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (San Jose, California, USA) (CHI ’16). Association for Computing Machinery, New York, NY, USA, 4647–4657. https://doi.org/10.1145/2858036.2858535

[10] Hao Fei, Yue Zhang, Yafeng Ren, and Donghong Ji. 2020. Latent Emotion Memory for Multi-Label Emotion Classification. Proceedings of the AAAI Conference on Artificial Intelligence 34, 05 (Apr. 2020), 7892–7899. https://doi.org/10.1609/aaai.v34i05.6271

[11] Ruidan He, Wue Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2017. An Unsupervised Neural Attention Model for Aspect Extraction. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics Volume 1: Long Papers. Association for Computational Linguistics, Vancouver, Canada, 388–397. https://doi.org/10.18653/v1/P17-1036

[12] Chenyang Huang, Amine Trabelsi, Xuebin Qin, Nawadh Farrague, and Omar R. Zaidan. 2019. SeqEmo: Multi-Label Emotion Classification Based on Latent Variable Chains Transformation. arXiv:1911.02147 [cs.CL]

[13] Mohammed Jabrell and Antonio Moreno. 2019. A Deep Learning-Based Approach for Multi-Label Emotion Classification in Tweets. Applied Sciences 9, 6 (2019). https://doi.org/10.3390/app9061123

[14] Pratik Kayal, Mayank Singh, and Pawan Goyal. 2020. Weakly-Supervised Deep Learning for Domain Invariant Sentiment Classification. In Proceedings of the 7th ACM IKDD-CDs and 25th COMAD (Hyderabad, India). COMAD COMAD 2020 Association for Computing Machinery, New York, NY, USA, 239–243. https://doi.org/10.1145/3371158.3371194

[15] 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:1907.11692 [cs.CL] http://arxiv.org/abs/1907.11692

[16] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In International Conference on Learning Representations. https://openreview.net/forum?id=BygkR6Cq37

[17] Saff 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. Association for Computational Linguistics, New Orleans, Louisiana, 1–17. https://doi.org/10.18653/v1/S18-1001

[18] Saff 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). European Language Resources Association (ELRA), Miyazaki, Japan. https://www.aclweb.org/anthology/L18-1030