All-in-One: Emotion, Sentiment and Intensity Prediction using a Multi-task Ensemble Framework

Md Shad Akhtar[1], Deepanway Ghosal[2†], Asif Ekbal[1], Pushpak Bhattacharyya[1], Sadao Kurohashi[3]
[1]*Department of Computer Science & Engineering, Indian Institute of Technology Patna, India.*
{shad.pcs15, asif, pb}@iitp.ac.in
[2]*School of Computer Science and Engineering, Nanyang Technological University, Singapore.*
dghosal@ntu.edu.sg
[3]*Department of Intelligence Science and Technology, Kyoto University, Japan.*
kuro@i.kyoto-u.ac.jp

***Abstract*—We propose a multi-task ensemble framework that jointly learns multiple related problems. The ensemble model aims to leverage the learned representations of three deep learning models (i.e., CNN, LSTM and GRU) and a hand-crafted feature representation for the predictions. Through multi-task framework, we address four problems of emotion and sentiment analysis, i.e., "emotion *classification* & *intensity*", "*valence*, *arousal* & *dominance* for emotion", "*valence* & *arousal* for sentiment", and "*3-class categorical* & *5-class ordinal classification* for sentiment". The underlying problems cover two granularity (i.e., *coarse-grained* and *fine-grained*) and a diverse range of domains (i.e., *tweets*, *Facebook posts*, *news headlines*, *blogs*, *letters* etc.). Experimental results suggest that the proposed multi-task framework outperforms the single-task frameworks in all experiments.**

***Index Terms*—Emotion Analysis, Sentiment Analysis, Intensity Prediction, Valence Prediction, Arousal Prediction, Dominance Prediction, Coarse-grained Emotion Analysis, Fine-grained Emotion Analysis, Fine-grained Sentiment Analysis, Multi-Layer Perceptron, Ensemble**

I. Introduction

Emotion analysis [1] deals with the automatic extraction of emotions expressed in a user written text. Ekman [2] categorized the basic human emotion as *anger*, *disgust*, *fear*, *surprise*, *sadness* and *joy*. In comparison, sentiment analysis [3] tries to automatically extract the subjective information from a user-written textual content and classify it into one of the predefined set of categories, e.g., *positive*, *negative*, *neutral* or *conflict*. Emotion [1] and sentiment [3] are closely related and are often been used incorrectly in a similar sense. According to Munezero et al. [4], emotions and sentiments differ on the scale of duration on which they are experienced. Emotions are usually shorter in duration, whereas sentiments are more stable and valid for longer period of time [5]. Also, sentiments are normally expressed towards a target entity, whereas emotions are not always target-centric [6]. For example, someone may wake up with joy without any valid reason.

These have applications in a diverse set of real-world problems such as stock market predictions, disaster management systems, health management systems, feedback systems for an organization or individual user to take an informed decision [7], [8], [9]. Any organization does not wish to lose their valuable customers. They can keep track of the emotions and sentiments of their customers over a period of time. If the unpleasant emotions or sentiments are being expressed by a customer on a regular basis, the organization can act in a timely manner to address his/her concerns. On the other hand, if the emotions and sentiments are pleasant, the organization can ride on the positive feedback of their customers to analyze and forecast their economic situation with more confidence.

The classification of emotions and sentiments into coarse-grained classes does not always reflect exact state of opinion of a user, hence, do not serve the purpose completely. Recently, the attention has been shifted towards fine-grained analysis on the dimensional scale [10], [11], [12], [13], [14]. Arousal or intensity defines the degree of emotion and sentiment felt by the user and often differs on a case-to-case basis. Within a single class (e.g., *Sadness*) some emotions are gentle (e.g., '*I lost my favorite pen today.*') while others can be severe (e.g., '*my uncle died from cancer today...RIP*'). Similarly, some sentiments are gentler than others within the same polarity, e.g., '*happy to see you again*' v/s '*can't wait to see you again*'.

The goal of the current study is to simultaneously solve four problems: (1) coarse-grained (categorical) emotion classification, (2) fine-grained (valence, arousal and dominance) emotion prediction, (3) fine-grained (valence and arousal) sentiment prediction, and (4) coarse-grained (categorical and ordinal) sentiment classification. We perform this by proposing an efficient multi-task ensemble framework.

Multi-task learning framework targets to achieve generalization by leveraging the inter-relatedness of multiple problems/tasks [15]. The intuition behind multi-task learning is that if two or more tasks are correlated then the joint-model can learn effectively from the shared representations. In comparison to single-task framework, where different tasks are solved in isolation, a multi-task framework offers three main advantages: (1) achieves better generalization; (2) improves the performance of each task through shared representation; and (3) requires only one unified model in contrast to separate models for each task in single-task setting, resulting in reduced

[†] The work was carried out while he was an undergraduate at IIT Patna.

complexity in terms of learnable model parameters.

Our proposed multi-task framework is greatly inspired from this, and it jointly performs multiple tasks. Our framework is based on an ensemble technique. At first, we learn hidden representations through three deep learning models, i.e., Convolutional Neural Network (CNN) [16], Long Short-Term Memory (LSTM) [17] and Gated Recurrent Unit (GRU) [18]. We subsequently feed the learned representations of three deep learning systems along with a hand-crafted feature vector to a Multi-Layer Perceptron (MLP) network to construct an ensemble. The objective is to leverage four different representations and capture the relevant features among them for predictions. The proposed network aims at predicting multiple outputs from the input representations in one-shot. We evaluate the proposed approach for four problems, i.e., *coarse-grained emotion analysis*, *fine-grained emotion analysis*, *fine-grained sentiment analysis*, and *coarse-grained sentiment analysis*. For *coarse-grained emotion analysis*, we aim to predict emotion class and its intensity value as the two tasks. The first task (i.e., *emotion classification*) classifies the incoming tweet into one of the predefined classes (e.g., *joy*, *anger*, *sadness*, *fear* etc.), while the second task (i.e., *emotion intensity prediction*) predicts the associated degree of emotion expressed by the writer in a continuous range of 0 to 1. In *fine-grained emotion analysis*, we aim to predict the *valence*, *arousal* and *dominance* scores in parallel, whereas, in the third problem, i.e., *fine-grained sentiment analysis*, our goal is to predict *valence* and *arousal* scores in a multi-task framework. The range of each task of the second and third problems is on the continuous scale of *1* to *5* (EmoBank [11]) and *1* to *9* (Facebook posts [12]), respectively. For the last problem, i.e., *coarse-grained sentiment analysis*, we solve message-level and topic-level sentiment prediction tasks together. The message-level task is a 3-class (positive, neutral and negative) categorical classification, while the topic-level task is a 5-class (highly positive (+2), positive (+1), neutral (0), negative (-1), and highly negative (-2)) ordinal classification. In total, we apply the proposed multi-task approach for four configurations: a) multi-task learning for classification (*emotion classification*) and regression (*emotion intensity prediction*) together; b) multi-task learning for two regression tasks together (sentiment *valence* & *arousal* prediction); c) multi-task learning for three regression tasks together (emotion *valence*, *arousal* & *dominance*); and d) multi-task learning for categorical (*message-level*) and ordinal (*topic-level*) classification tasks.

The main contributions of our proposed work are summarized below: **a)** we effectively combine deep learning representations with manual features via an ensemble framework; and **b)** we develop a multi-task learning framework which attains overall better performance for different tasks related to emotion, sentiment and intensity.

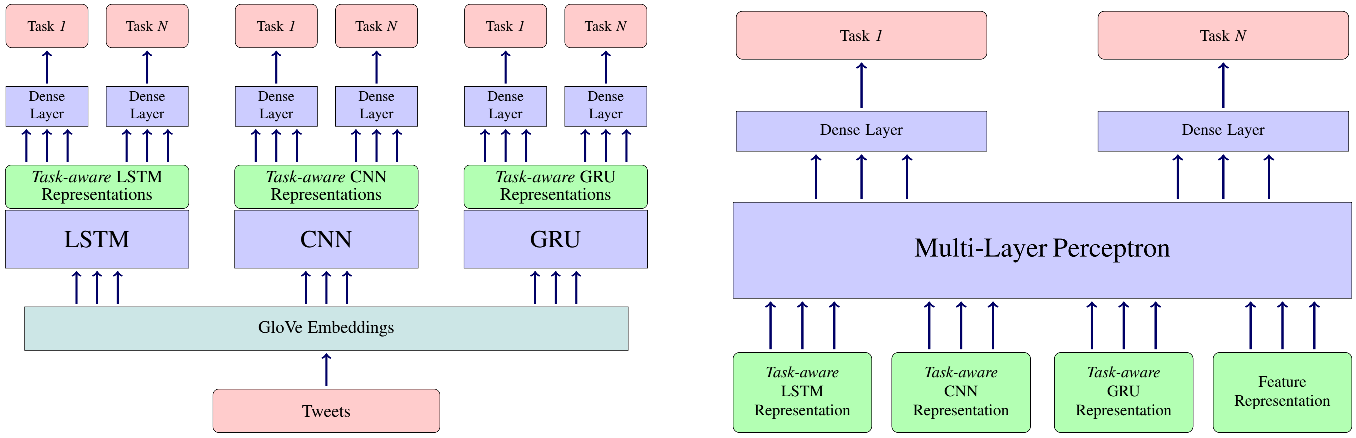
The rest of the paper is organized as follows. In Section II we present a brief overview of the related works. Section III describes our proposed methodology in details. In Section IV, we present our datasets, experimental setup, results along with necessary analysis. Finally, we conclude in Section V with future research directions.

II. Related Work

Literature suggests that multi-task learning has been successfully applied in a multitude of machine learning (including natural language processing) problems [19], [20], [21], [22]. Authors in [21] employed recurrent neural network for their multi-task framework where they treated *3-way* classification and *5-way* classification as two separate tasks for sentiment analysis. An application of convolutional neural network (CNN) in multi-task framework has been proposed in [19] for predicting multiple Natural Language Processing (NLP) tasks, e.g., Part-of-Speech (PoS) tagging, Chunking, Named Entity Recognition (NER) etc. Similarly, authors in [20] adopted deep multi-task learning for solving many NLP problems. They showed that supervising low-level tasks at lower levels of deep network can improve the performance of higher-level tasks. Lu et al. [23] proposed a multi-task model that leverages the weight-shared parameters for learning the representation of text in neural network model for sentiment classification. In another work, Fraisse and Paroubek [24] developed a dataset, namely OSE, for the unified modelling of opinion, sentiment and emotion. They categorized the subjective information into three broad groups, i.e., *intellective*, *affective-intellective*, and *affective* expressions representing the expressed opinion, sentiment, and emotion of the user.

One of the earlier works on emotion detection looks at emotion bearing words in the text for classification [25]. In another work, Dung et al. [26] studied human mental states with respect to an emotion for training a Hidden Markov Model (HMM). In contrast, authors in [27] proposed a rule-based approach to extract emotion-specific semantics, which is then utilized for learning through various separable mixture models. These systems concentrated on emotion classification, whereas, the works reported in [13], [28], [29] focus only on intensity prediction. Jain et al. [28] used an ensemble of five different neural network models for predicting the emotion intensity. They also explored the idea of multi-task learning in one of the models, where they treated four different emotions as the four tasks. The final predictions were generated by a weighted average of the base models. Koper et al. [29] employed a random forest regression model on the concatenated lexicon features and Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM) features. Authors in [30] employed LSTM and SVR in cascade for predicting the emotion intensity. In another work, [22] have proposed VA (*Valence-Activation*) model for emotion recognition in 2D continuous space. Recently, Xu et al. [31] proposed an emotion-aware embeddings (Emo2Vec) that encodes the emotional semantics into vector through the application of multi-task learning. Further, they established that the emotion-aware embeddings, on an average, has better performance than various existing embeddings (e.g., GloVe [32], Sentiment-Specific [33] & DeepMoji[34] embeddings) across multiple emotion and sentiment related tasks.

Following the trends of emotion intensity prediction, researchers have also focused on predicting the intensity score for sentiment [35], [36], [37], [38]. Balahur et al [35] studied the behavior of sentiment intensity for text summarization.



(a) Individual multi-task deep learning models

(b) Multi-task ensemble model

Fig. 1: Proposed Multi-task framework.

Authors in [37] proposed semi-supervised technique that used sentiment bearing word embeddings to produce a continuous ranking among adjectives that share common semantics. In another work [38], authors proposed a stacked ensemble technique for sentiment intensity prediction in financial domain. Traditional techniques, e.g., boosting [39], bagging [40], voting (weighted, majority) [41] etc. are some of the common choices for constructing ensemble [42], [43], [44]. Akhtar et al. [45] proposed an ensemble technique based on Particle Swarm Optimization (PSO) to solve the problem of aspect based sentiment analysis.

Literature suggests that the intermediate-level representation - learned on a task - has been successfully utilized for learning another task [46]. However, our work differs from [46] in the sense that we aim to combine different intermediate-level representations through an ensemble network for the same tasks in a multi-task learning framework. Further, our proposed approach differs from these existing systems in terms of the following aspects: a) our MLP based ensemble addresses both *classification* and *regression* problems; b) our multi-task framework handles a diverse set of tasks (i.e., *classification & regression problems, two regression problems*, *three regression problems*, and *categorical & ordinal classification problems*); and c) our proposed approach covers two granularity (i.e., coarse-grained & fine-grained) and a diverse set of domains (i.e., *tweets, Facebook posts, news headlines, blogs* etc.).

III. Proposed Methodology

Ensemble is an efficient technique in combining the outputs of various candidate systems. The basic idea is to leverage the goodness of several systems to improve the overall performance. Ensemble solves three important machine learning issues, *viz.* statistical, computational and representational [47]. Statistical issues arise in absence of sufficient training data and each individual system has different hypothesis. Ensemble aims to find an accurate hypothesis by averaging hypotheses of individual systems. Computational issue arises when participating systems stuck at a local minimum. Through ensemble the search for a hypothesis can start at a different point of the search space, thus minimize the possibility of getting stuck at the local minimum. The third issue is representational which arises when none of participating hypotheses can approximate the training data. In such case a hypothesis can be approximated through the weighted sum of various hypotheses.

Most of the existing ensemble methods [43], [44], [42], [45] addressed the classification problem, whereas our proposed ensemble technique is developed to solve both the classification and regression problems at the same time. In addition, our problem domain differs from these existing systems.

Motivated by this, we propose a multi-task ensemble learning framework built on top of learned representations of three deep learning models and a hand-crafted feature vector. We separately train all three deep learning models, i.e., a CNN, a LSTM and a GRU network in a multi-task framework (Figure 1a). Once the network is trained, we extract an intermediate layer activation from these CNN, LSTM and GRU models.These three *task-aware* deep representations are concatenated with a feature vector before feeding into the multi-task ensemble model. The multi-task ensemble model is a MLP network which comprises of four hidden layers. The first two hidden layers are shared for all the tasks (i.e., the hidden representation jointly captures the relationship of all the input representations) and the final two hidden layers are specific for each individual task to learn the mapping of the shared hidden representation and output labels. The idea is to exploit the richness of different feature representations and to learn a combined representation for solving multiple tasks.

We tried with different number of hidden layers (i.e., 2 layers (1 shared + 1 task-specific), 3 layers (2 shared + 1 task-specific), 4 layers (2 shared + 2 task-specific) and 5 layers (3 shared + 2 task-specific)) for the ensemble network and observe better performance with 4 layers. Consequently, we show that the ensemble model performs better than each of the individual models. A high-level outline of the proposed approach is depicted in Figure 1. Figure 1a shows the multi-task framework for the individual CNN, LSTM and GRU models. After training, the respective *task-aware* intermediate representations (*color coded green* in Figure 1a) and the hand-crafted feature vector are used as input for the ensemble in Figure 1b.

A. Deep Learning Models

We employ the architecture of Figure 1a to train and tune all the deep learning models using pre-trained GloVe (common crawl 840 billion) word embeddings [32]. In our CNN model, we use two convolution followed by max-pool layers (*conv-pool-conv-pool*). Each convolution layer has 100 filters sliding over 2, 3 and 4 words in parallel. For LSTM/GRU models, we use two stacked LSTM/GRU layers, each having 128 neurons. The CNN, LSTM and GRU layers are followed by two task-specific fully connected layers and the output layer. We use 128 (*color coded green* in Figure 1a) and 100 (*color coded blue* 'Fully-connected Layer' in Figure 1a) neurons in the fully connected layers for all the models. The output layer has multiple neurons depending on the number of tasks in the multi-task framework. The fully connected layer activation is set to *rectified linear* [48], and the output layer activation is set according to the task - *softmax* for classification & *sigmoid* for regression. We apply 25% *Dropout* [49] in the fully-connected layers as a measure of regularization. The *Adam* [50] optimizer with default parameters is used for gradient based training. It should be noted that we use the same hyper-parameters for all the models in order to maintain consistency. During validation phase, we have experimented with different network configurations (e.g., number of hidden layers/units, activation functions, dropout etc.). For dropout, we experimented with the different values ranging from 0.1 - 0.6. Finally, we have chosen the parameter configuration that is well-suited to all the datasets (e.g., $dropout = 0.25$ in our case). We present the details of hyper-parameters for training of the neural networks in Table I.

Parameter	**EmoInt - 2017**	**EmoBank**	**FB post**	**SemEval-2016**
Loss	Emotion class. - *Cross-Ent* Emotion intensity - *MSE*	*MSE*		*Cross-Ent*
Hidden activations	*ReLU* [48]			
Output activations	Emotion class. - *Softmax* Emotion intensity - *Sigmoid*	*Sigmoid*		*Softmax*
Shared Layers	*CNN* - 2 (conv-pool-conv-pool) *LSTM* - 2 (128 neurons each) *GRU* - 2 (128 neurons each) *MLP (Feat)* - 1 (128 neurons) *Ensemble* - 2 (128 & 100 neurons)			
Task-specific FC Layers	*Base models* - 2 (128 & 100 neurons) *Ensemble* - 2 (64 & 32 neurons)			
Convolution filter	*100* filters of size *2*, *3* & *4* in parallel			
Batch	*64*			
Epochs	*40*			
Dropout [51]	*25%*			
Optimizer	Adam [50]			

TABLE I: Various hyper-parameters for training of the deep learning models.

B. Hand-Crafted Feature Vector

In addition to the deep learning representations we extract and use the following set of features for constructing the ensemble model.

- **Word and Character Tf-Idf:** Word Tf-Idf weighted counts of 1, 2, 3 grams and character Tf-Idf weighted counts of 3, 4 and 5 grams.
- **TF-Idf Weighted Word Vector Averaging:** Word embeddings models are generally good at capturing semantic information of a word. However, every word is not equally significant for a specific problem. Tf-Idf assigns weights to the words according to their significance in the document. We scale the embeddings of words in the text according to their Tf-Idf weights and use this average of weighted embeddings of words to create a set of features.
- **Lexicon Features:**
  - count of positive and negative words using the MPQA subjectivity lexicon [52] and Bing Liu lexicon [53].
  - positive, negative scores from Sentiment140, Hashtag Sentiment lexicon [54], AFINN [55] and Sentiwordnet [56].
  - aggregate scores of hashtags from NRC Hashtag Sentiment lexicon [54].
  - count of the number of words matching each emotion from the NRC Word-Emotion Association Lexicon [57].
  - sum of emotion associations in NRC-10 Expanded lexicon [58], Hashtag Emotion Association Lexicon [59] and NRC Word-Emotion Association Lexicon [57].
  - Positive and negative scores of the emoticons obtained from the AFINN project [55].
- **Vader Sentiment:** We use Vader sentiment [60] which generates a compound sentiment score for a sentence between -1 (extreme negative) and +1 (extreme positive). It also produces ratio of positive, negative and neutral tokens in the sentence. We use the score and the three ratios as features in our feature based model.

Since the feature vector dimension is too large in comparison with DL representation during ensemble, we project the feature vector to smaller dimension (i.e., 128) through a small MLP network.

IV. Datasets, Experiments and Analysis

A. Datasets

We evaluate our proposed model on the benchmark datasets of 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA-2017) shared task on emotion intensity (EmoInt-2017) [13], EmoBank [11], SemEval-2016 shared task on Sentiment Analysis in Twitter [61], and Facebook posts [12] for the *coarse-grained emotion analysis*, *fine-grained emotion analysis*, *coarse-grained sentiment analysis*, and *fine-grained sentiment analysis*, respectively.

Datasets		**Train**	**Validation**	**Test**	**Total**
EmoInt [13]	Anger	857	84	760	1,701
	Fear	1147	110	995	2,252
	Sad	786	74	673	1,533
	Joy	823	79	714	1,616
	Total	3,613	347	3,142	7,102
EmoBank [11]		7044 (70%)	1006 (10%)	2012 (20%)	10,062
Facebook Post [12]		*10-fold cross-validation*			2895
SemEval-2016 [61]		6000	2000	20,632	28,632

TABLE II: Dataset statistics.

Text	**Emotion**	**Intensity**
Just died from laughter after seeing that.	Joy	0.92
My uncle died from cancer today...RIP.	Sadness	0.87
Still salty about that fire alarm at 2am this morning.	Fear	0.50
Happiness is the best revenge.	Anger	0.25

(a) **Coarse-grained emotion analysis**: Intensity is on the scale of 0 to 1 [13].

Text	**Valence**	**Arousal**	**Dominance**
I am thrilled with the price.	4.4	4.4	4.0
I hate it, despise it, abhor it!	1.0	4.4	2.2
I was feeling calm and private that night.	3.2	1.6	3.0
I just hope they keep me here.	2.7	2.7	2.0
James Brown's 5-year-old son left out of will.	1.0	2.6	2.2
I shivered as I walked past the pale mans blank eyes, wondering what they were staring at.	1.2	3.0	1.5

(b) **Fine-grained emotion analysis**: Valence, arousal & dominance are on the scale of 1 to 5 [11].

Text	**Valence**	**Arousal**
I bought my wedding dress Monday and I cant wait to have it on again!!!! its sooo beautiful.	8.0	8.0
Happy, got new friends, and lifes getting smoother.	8.0	1.5
At least 15 dead as ###### forces attack &&&& aid ships!!!!!!! i hhhhhhate ######	1.5	8.0
The worst way to miss someone is when they r right beside u and yet u know u can never have them.	2.5	1.5

(c) **Fine-grained sentiment analysis**: Valence and arousal are on the scale of 1 to 9 [12].

TABLE III: Multi-task examples of emotion analysis and sentiment analysis from benchmark datasets. ***Valence*** ⇒ Concept of polarity (pleasant / unpleasant); ***Arousal or Intensity*** ⇒ Degree of emotion/sentiments; ***Dominance*** ⇒ Control over a situation;

The dataset of EmoInt-2017 [13] contains generic tweets representing four emotions, i.e., *anger*, *fear*, *joy* and *sadness* and their respective intensity scores. It contains 3613, 347 & 3142 generic tweets for training, validation and testing, respectively. The EmoBank dataset [11] comprises of 10,062 tweets across multiple domains (e.g., *blogs*, *new headlines*, *fiction* etc.). Each tweet has three scores representing *valence*, *arousal* and *dominance* of emotion concerning the writer's and reader's perspective. Each score has continuous range of *1* to *5*. For experiments, we adopt 70-10-20 split for training, validation and testing, respectively. The Facebook posts dataset [12] has 2895 social media posts. Posts are annotated on a nine-point scale with valence and arousal score for sentiment analysis by two psychologically trained annotators. We perform *10-fold* cross-validation for the evaluation. The SemEval-2016 [61] dataset contains approximately 28K tweets for message and topic level sentiment analysis. These messages are distributed over 60, 20 and 100 different topics in training, validation, and test datasets, respectively. In message-level task, each message (or tweet) is labelled as either *positive*, *negative* or *neutral*, whereas, in topic-level task, each message with respect to a given topic has ordinal classification as *highly positive*, *positive*, *neutral*, *negative*, and *highly negative*. A summary of the datasets statistics are depicted in Table II.

Few example scenarios for the problems of emotion analysis (coarse-grained & fine-grained) and sentiment analysis (fine-grained) are depicted in Table III. In the first example shown in Table IIIa, emotion *'joy'* is derived from the phrase '*died from laughter*' which is intense. However, the emotion associated with the second example which contains similar phrase '*died from cancer*' is *'sadness'*. The third example expresses *'fear'* with mild intensity, whereas, the fourth example conveys *'anger'* emotion with relatively lesser intensity. Similarly, examples of fine-grained emotion analysis are listed in Table IIIb. Each text is associated with psychologically motivated VAD (*Valence*, *Arousal* & *Dominance*) scores. *Valence* is defined by pleasantness (positive) or unpleasantness (negative) of the situations. *Arousal* reflects the degree of emotion, whereas, *Dominance* suggests the degree of control over a particular situation. Similarly, Table IIIc depicts the example scenarios for fine-grained sentiment analysis.

B. Experimental Setup and Results

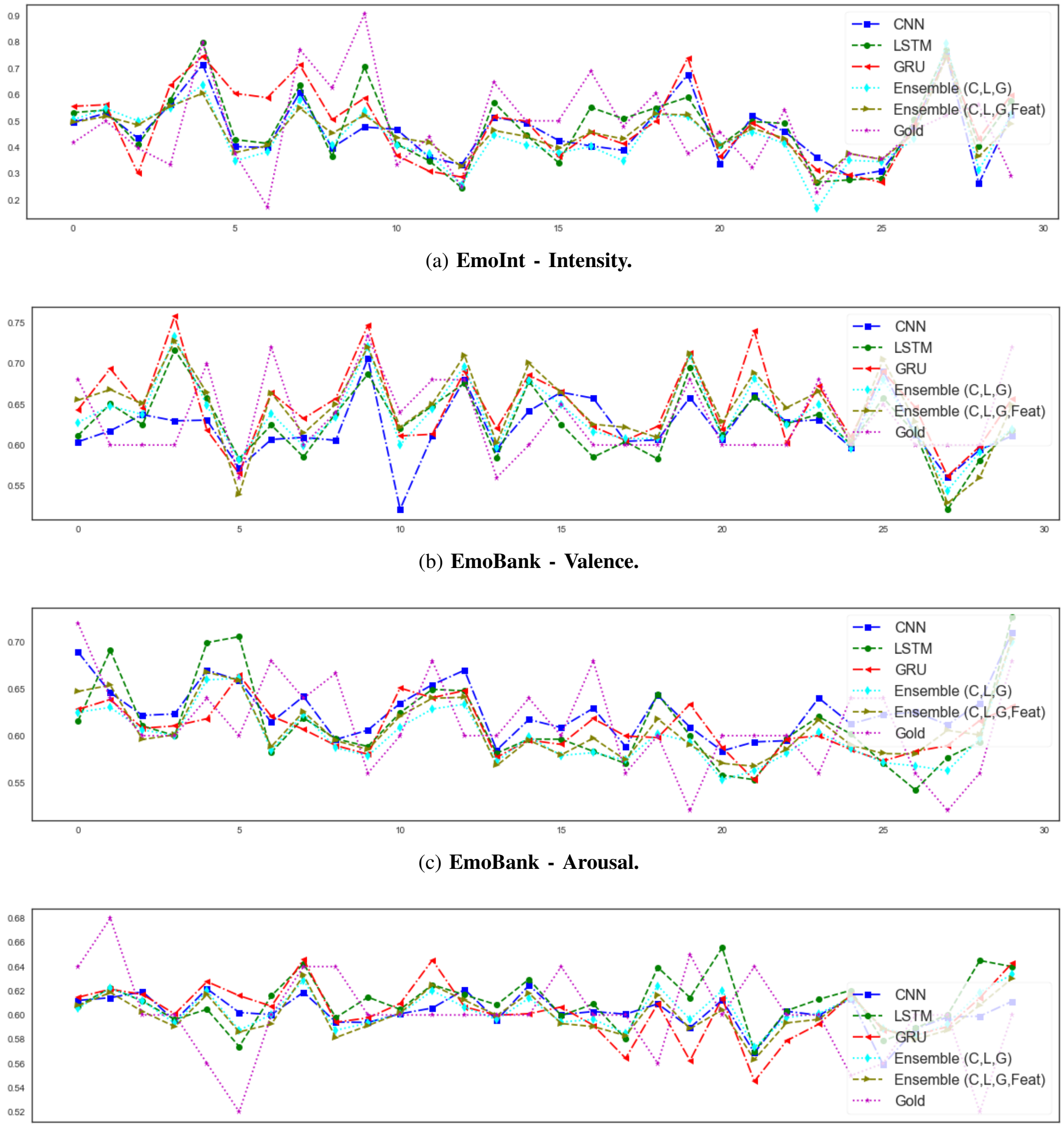
We use Python based libraries, Keras [62] and Scikit-learn [63] for implementation. For evaluation, we compute *accuracy* for the classification (*emotion class*) and *pearson correlation coefficient* for the regression (e.g., *intensity*, *valence*, *arousal* & *dominance*). Pearson correlation coefficient measures the linear correlation between the actual and predicted scores. The choice of these metrics was inspired from [13] and [12]. We normalize the *valence*, *arousal* and *dominance* scores on a *0* to *1* scale. For prediction, we use *softmax* for classification and *sigmoid* for regression.

Table IV shows the results on the test set for coarse-grained emotion analysis. In multi-task framework, we predict the emotion class and intensity together, whereas in single-task framework we build two separate models, one for classification and one for intensity prediction. We follow a dependent evaluation[1] technique where we compute the scores of only those instances which are correctly predicted by the emotion classifier. Such evaluation is informative and realistic as predicting intensity scores for the misclassified instances would

[1] Please note that we adopted dependent evaluation strategy as this is commonly used for the evaluation of related-tasks in multi-task framework.

Models	Multi-task learning		Single-task learning	
	Emotion Class	Intensity*	Emotion Class	Intensity*
	Accuracy %	Pearson	Accuracy %	Pearson
CNN (C)	80.52	0.578	79.56	0.493
LSTM (L)	84.69	0.625	84.02	0.572
GRU (G)	84.94	0.606	83.45	0.522
Feat (MLP)	78.32	0.576	78.10	0.572
Ensemble (C, L & G)	85.93	0.657	85.77	0.596
Ensemble (C, L, G & Feat)	**89.88**	**0.670**	89.52	0.603
Ensemble (C, L, G & Feat) - End2End	82.46	0.604	75.65	0.549
Significance T-test (p-values)	*0.008*	*0.040*	-	-

TABLE IV: **Coarse-grained Emotion Analysis:** Experimental results for multi-task (*i.e. single model for both tasks in parallel*) and single-task (*i.e. first a tweet is classified to an emotion class and then intensity is predicted only for the correctly classified tweets*) learning framework for **EmoInt-2017 datasets** [13]. ***We evaluate emotion intensity only for those instances whose respective class was correctly predicted**.



(a) **EmoInt - Intensity.**

(b) **EmoBank - Valence.**

(c) **EmoBank - Arousal.**

(d) **EmoBank - Dominance.**

Fig. 2: Contrasting nature of the individual models and improved scores after ensemble for emotion intensity prediction. **X-axis**: 30 random samples from the test set. **Y-axis**: Intensity values.

Models	Multi-task learning			Single-task learning		
	Valence	Arousal	Dominance	Valence	Arousal	Dominance
CNN (C)	0.567	0.347	0.234	0.552	0.334	0.222
LSTM (L)	0.601	0.337	0.245	0.572	0.318	0.227
GRU (G)	0.569	0.315	0.243	0.553	0.306	0.227
Feat (MLP)	0.600	0.324	0.248	0.590	0.320	0.223
Ensemble (C, L & G)	0.618	0.365	0.263	0.603	0.351	0.234
Ensemble (C, L, G & Feat)	**0.635**	**0.375**	**0.277**	0.616	0.355	0.237
Ensemble (C, L, G & Feat) - End2End	0.594	0.317	0.218	0.589	0.289	0.207
Significance T-test (p-values)	*0.048*	*0.027*	*0.310*	-	-	-

TABLE V: **Fine-grained Emotion Analysis:** Experimental results (Pearson correlation) for multi-task and single-task learning framework on **EmoBank datasets** [11].

Models	Multi-task learning		Single-task learning	
	Valence	Arousal	Valence	Arousal
CNN (C)	0.678	0.290	0.666	0.283
LSTM (L)	0.671	0.324	0.655	0.315
GRU (G)	0.668	0.313	0.657	0.294
Feat (MLP)	0.672	0.291	0.671	0.259
Ensemble (C, L & G)	0.695	0.336	0.684	0.324
Ensemble (C, L, G & Feat)	**0.727**	**0.355**	0.713	0.339
Ensemble (C, L, G & Feat) - End2End	0.722	0.313	0.713	0.303
Significance T-test (p-values)	*0.033*	*0.024*	-	-

TABLE VI: **Fine-grained Sentiment Analysis:** Experimental results (Pearson correlation) for multi-task and single-task learning framework on **FB post datasets** [12].

Models	Multi-task learning		Single-task learning	
	Message-level	Topic-level	Message-level	Topic-level
	Accuracy	MAE	Accuracy	MAE
CNN (C)	52.28	1.03	50.14	1.40
LSTM (L)	53.55	1.11	51.06	1.40
GRU (G)	54.21	0.94	53.55	1.12
Feat (MLP)	56.11	0.99	53.38	1.01
Ensemble (C, L & G)	56.65	0.92	56.03	0.94
Ensemble (C, L, G & Feat)	**57.11**	**0.91**	55.60	0.92
Ensemble (C, L, G & Feat) - End2End	50.99	1.03	49.91	1.09
Significance T-test (p-values)	*0.129*	*0.277*	-	-

TABLE VII: **Coarse-grained Sentiment Analysis:** Experimental results for multi-task and single-task learning framework on **SemEval-2016 datasets** [61].

not convey the correct information. For direct comparison, we also adopted a similar approach for intensity prediction evaluation in the single-task framework. The first half of Table IV reports the evaluation results for three deep learning and one feature-driven models. In multi-task framework, CNN reports 80.52% accuracy for classification and 0.578 Pearson score for intensity prediction. The multi-task LSTM and GRU models obtain 84.69% & 84.94% accuracy values and 0.625 & 0.606 Pearson scores, respectively. The hand-crafted features when subjected to a MLP network yields 78.32% accuracy and 0.576 Pearson score. The corresponding models in single-task framework report 79.56%, 84.02%, 83.45% & 78.10% accuracy values and 0.493, 0.572, 0.522 & 0.572 Pearson scores for CNN, LSTM, GRU & feature-based models, respectively. It is evident that multi-task models perform better than the single-task models by a convincingly good margin for intensity prediction, and better for class prediction. On further analysis, we observe that these models obtain quite similar performance numerically. However, they are quite contrasting on a qualitative side. Figure 2 shows the contrasting nature of different individual models for emotion intensity. In some cases, prediction of one model is closer to the gold intensity than the other models and vice-versa. An ensemble system constructed using only deep learning models achieves the enhanced accuracy of 85.93% and Pearson score of 0.657. Further inclusion of hand-crafted feature vectors (c.f. section III-B) in the ensemble network results in an improvement of around 4% accuracy and 1.5% Pearson score.

We report the results for *fine-grained emotion analysis*, *fine-grained sentiment analysis*, and *coarse-grained sentiment*

Models	Multi-task learning		Single-task learning	
	Valence(EmoBank)	Valence(FB post)	Valence(EmoBank)	Valence(FB post)
CNN (C)	0.561	0.703	0.552	0.666
LSTM (L)	0.580	0.689	0.572	0.655
GRU (G)	0.558	0.686	0.553	0.657
Feat (MLP)	0.604	0.714	0.590	0.671
Ensemble (C, L & G)	0.603	0.708	0.603	0.684
Ensemble (C, L, G & Feat)	**0.625**	**0.730**	0.616	0.713
Ensemble (C, L, G & Feat) - End2End	0.604	0.722	0.589	0.713

TABLE VIII: Experimental results for two different datasets (**EmoBank-Valence** [11] and **FB post-Valence** [12]) jointly trained in the multi-task learning framework. The FB post-Valence leverages the availability of larger EmoBank dataset for the performance improvement (c.f. Table VI).

analysis in Tables V, VI and VII, respectively. Similar to *coarse-grained emotion analysis*, we observe that multi-task models in *fine-grained emotion analysis* achieve the improved Pearson scores (0.635, 0.375 & 0.277) as compared to the single-task models (0.616, 0.355 & 0.237) for the three tasks, i.e., valence, arousal and dominance, respectively. The ensemble approach also achieves better performance compared to each of the base models for all the tasks. For fine-grained sentiment analysis, deep learning based models, i.e., CNN, LSTM, GRU & MLP (Feat) obtain the Pearson scores of 0.678, 0.671, 0.668 & 0.672, respectively, for *valence* in multi-task environment. The ensemble of these DL models and handcrafted feature representation via MLP obtains an increased Pearson score of 0.727. The proposed approach also achieves the best Pearson score of 0.355 for *arousal*.

One of the important use-case of multi-task learning is the effective utilization of a larger dataset for the performance improvement of a dataset with insufficient training samples. For all four datasets, we extract two related information from a single input in a multi-task framework. Therefore, for the completeness, we also exploit the joint-learning of two related tasks for two different datasets in the proposed multi-task framework, i.e., valence prediction for emotion (EmoBank) and valence prediction for sentiment (FB post). As mentioned in Table II, out of the two datasets, EmoBank dataset [11] has relatively more number of samples than the FB post dataset [12]. We hypothesis that the valence prediction for sentiment would leverage the increase in training samples. We, at first, train our proposed multi-task network for the emotion valence prediction, and then, further, train the network for the sentiment valence prediction. We report our experimental results in Table VIII. We observe that the sentiment valence prediction, indeed, leverages the availability of the larger dataset for the performance enhancement over the sentiment valence and arousal case in fine-grained emotion analysis (c.f. Table VI). Since the emotion valence prediction does not have such advantage, we observe that it has comparable performance with the emotion valence, arousal and dominance case in fine-grained sentiment analysis (c.f. Table V).

We also evaluate our proposed model in an end-to-end network, where all the base models and the ensemble model are trained in an unified architecture. We list the obtained results in Tables IV, V, VI, and VII for *coarse-grained emotion*, *fine-grained emotion*, *fine-grained sentiment* and *coarse-grained sentiment*, respectively. It is evident form these results that the multi-task learning performs better than the single-task learning framework in all cases. Further, we observe that the obtained results in an end-to-end network are inferior to the proposed approach where all the base models are trained and tuned separately, and the learned representations are employed for the final ensemble. This could be because the end-to-end model has a comparatively large set of parameters to learn, and since, the number of training samples in our datasets are not sufficient, the model finds its non-trivial to optimize the learnable parameters.

We observe two phenomena from our experimental results: **a)** use of multi-task framework for the related tasks indeed helps in achieving generalization; and **b)** the ensemble network leverages the learned representations of three base models & the feature vector and produces superior results. We also perform statistical significance test (*T-test*) on the 10 runs of the proposed approach. The *p*-values (reported in Table IV, V & VI) suggest that the results of MTL framework is statistically significant than STL framework with 95% confidence for all the datasets (except EmoBank-Dominance and SemEval-2016 datasets).

C. Comparative Analysis

For *coarse-grained sentiment analysis*, we compare our proposed approach with Prayas system [28], which was the top performing system at EmoInt-2017 [13] shared task on Emotion Intensity. Prayas [28] used an ensemble of five different neural network models including a multitasking feed-forward model. Although the final model was built for each emotion type separately, in multi-task model the authors treated four emotion classes as the four tasks. However, our proposed approach treats emotion classification and emotion intensity prediction as two separate tasks, and then learns jointly (a completely different setup than Prayas). Prayas reported the Pearson score of 0.662 for emotion intensity. In comparison, our proposed approach obtains a Pearson score of 0.670 for dependent evaluation, and 0.647 for independent evaluation. Statistical *T-test* shows that the value (0.670) is statistically significant over the model of Prayas.

Similarly, we do not compare our proposed approach with the other systems of EmoInt-2017 [13] because of the following two reasons: a) those systems are of single-task nature as compared to our proposed multi-task; and b) separate models

Models	Emotion Class		Emotion Intensity	
	Accuracy	F1-score	Dependent Evaluation	Independent Evaluation
Baseline+	-	-	-	0.648
Prayas (Multi-task)*	-	-	-	0.662
System [64]	-	88.00	-	-
Proposed (Single-task)	89.52	89.32	-	0.603
Proposed (Multi-task)	**89.88**	**89.73**	**0.670**	0.647

TABLE IX: **Coarse-grained Emotion Analysis:** Comparative results. **Dependent evaluation:** Intensity was evaluated only if its respective class was correctly predicted; **Independent evaluation:** Intensity score is evaluated independent of the emotion class; +Baseline system is taken from [13]. *****Prayas [28] was the top system at EmoInt-2017. Multi-task system of the proposed approach and Prayas are different. Prayas treated intensity prediction of four emotion classes as multi tasks, whereas, we addressed emotion classification and intensity prediction as two tasks.

Ensemble	EmoInt		EmoBank			FB post		SemEval	
	Class	Intensity	Valence	Arousal	Dominance	Valence	Arousal	Message-level	Topic-level
	Acc	Pearson	Pearson			Pearson		Acc	MAE
Proposed (MTL)	**89.58**	**0.670**	**0.635**	**0.375**	**0.277**	**0.727**	**0.355**	**57.11**	**0.91**
GradientBoost	82.59	0.545	0.596	0.292	0.219	0.705	0.273	54.88	1.91
AdaBoost	70.43	0.451	0.562	0.237	0.157	0.677	0.225	53.78	1.90
Bagging	65.21	0.492	0.550	0.213	0.166	0.681	0.222	52.62	1.88
Voting	77.24	0.540	0.592	0.292	0.213	0.704	0.293	56.17	1.94

TABLE X: Comparison with the traditional ensemble approaches.

Models	Valence	Arousal
	Pearson correlation	
System [12]	0.650	**0.850**
System - X*	0.390	0.105
Proposed (Single-task)	0.713	0.339
Proposed (Multi-task)	**0.727**	0.355

TABLE XI: **Fine-grained Sentiment Analysis:** Comparative results for Facebook posts dataset. **System - X***: *Google search lists this paper in the citation list of [12], however, the publication details are not available. The pdf is available at www.goo.gl/DcdaHF.*

were trained for each of the emotions and an average score was reported as compared to an unified single model that addressed all the emotions and their intensity values altogether. The baseline system for emotion intensity prediction in Table IX is taken from [13], which also differs from our proposed approach considering the above two points, and hence does not provide an ideal candidate for direct comparison.

For emotion classification, we compare our obtained results with [64]. Authors in [64] carried out an extensive study on various emotion datasets, and also performed both *in-corpus* and *cross-corpus* classification experiments. They reported F1-score of 88.00% for emotion classification for EmoInt-2017 dataset. In comparison, our STL and MTL frameworks report an increased F1-score of 89.32% and 89.73%, respectively. Further, the MTL result is statistically significant over [64] with p-value $= 0.0011$. A comparative analysis is presented in Table IX.

The datasets for *fine-grained emotion analysis* and *fine-grained sentiment analysis* problems, i.e., EmoBank [11] and Facebook posts [12] are relatively recent datasets and limited studies are available on these. We did not find any existing system that evaluated Pearson score for these datasets except the resource paper of Facebook posts [12]. For *valence* in *fine-grained sentiment analysis* a Pearson score of 0.650 has been reported in [12] using a Bag-of-Words (BoW) model. In comparison, our proposed approach reports the Pearson score of 0.727, an improvement of 7 points. For *arousal* Preoţiuc-Pietro et al. [12] reported a Pearson score of 0.850 as compared to 0.355 of ours. It should be noted that we tried to reproduce the scores of [12] using the same BoW model. We obtained the similar Pearson score of 0.645 for *valence*, however, we could not reproduce the reported results for arousal (we obtained Pearson score of 0.27)[2]. In Table XI, we demonstrate the comparative results for *fine-grained sentiment analysis*.

Furthermore, we compare our proposed ensemble framework with various traditional ensemble approaches such as Gradient Boost [65], AdaBoost [66], Bagging [67] and Voting [41]. The comparative results are reported in Table X for all four datasets. The obtained results suggest that the proposed ensemble approach yields much better predictions that any of the traditional ensemble approaches.

D. Error Analysis

For the emotion classification problem we analyze the confusion matrix and observe that the proposed model often confuses between *fear* and *sadness* class. In total 80 tweets ($\sim$8%) representing *fear* are misclassified as *sadness*, whereas, 40 instances ($\sim$6.5%) of *sadness* are misclassified as *fear*. The confusion matrix is depicted in Table XIII.

[2] Please note that the research reported in www.goo.gl/DcdaHF obtained only 0.105 on the same dataset

	Text	Actual	Predicted	Possible Reason
EmoInt-2017	**Emotion Classification**			
	*Going back to **blissful ignorance**.*	Sad	Joy	Metaphoric sentence.
	*I know if I was not an optimist I would **despair**.*	Fear	Sad	Strong expressions.
	*Class is canceled due to a **funeral**, not sure if it is appropriate to be happy or sad.*	Joy	Sad	
	Intensity Prediction			
	*Just **died from laughter** after seeing that.*	Joy/0.92	Joy/0.50	Metaphoric sentence.
	*Never let the sadness of your past **ruin** your future.*	Sad/0.29	Sad/0.64	Strong expression.
EmoBank	**Valence Prediction**			
	*News **Baby pandas! Baby pandas! Baby pandas!***	4.4	2.9	Repeated entities.
	*It's **summertime**, so it must be time for **CAMP!***	4.4	3.1	Implicit emotion.
	Arousal Prediction			
	*Carter was a **disaster**, said Ford.*	4.0	3.0	Implicit emotions.
	*The company is **on a roll.***	4.0	2.8	
	Dominance Prediction			
	*Three days later, another **B-29** from the **509th** bombed Nagasaki.*	2.0	3.3	Numerical entities.
	*The company reported a net loss of **$608,413** or **39** cents a share, compared with year-earlier net income of **$967,809** or **62** cents a share.*	2.2	3.4	
Facebook Posts	**Valence Prediction**			
	*I am on **cloud nine** right now.*	7.5	4.3	Idiomatic expressions.
	*We'll be **off and running** to a lil' place called SILVERWOOD today! Can't wait! :)*	9.0	4.8	
	Arousal Prediction			
	Thank you** all for **wishing** me a **happy birthday.	1.5	8.1	Strong expressions.
	***Happy turkey day** everybody.*	1.5	7.5	

TABLE XII: **Error Analysis:** Frequent error cases for the best performing multi-task models.

Sadness	15	9	40	548
Fear	16	17	901	80
Joy	11	657	25	12
Anger	718	31	29	33
	Anger	Joy	Fear	Sadness

TABLE XIII: Confusion matrix for EmoInt-2017 emotion classification problem.

We also perform qualitative error analysis on the predictions of our best performing multi-task models. At first, we identify the most commonly occurring errors and then we analyze 15 test instances for each such error to detect the common error patterns. A number of frequently occurring error cases along with their possible reasons are shown in Table XII. We observe that the main sources of errors are metaphoric sentences, strong expressions, implicit emotions and idiomatic expressions. For example, presence of metaphoric phrases "*blissful ignorance*" and "*died from laughter*" in the tweets seems to guide the system in incorrect predictions for emotion classification and intensity prediction, respectively.

We also compare the predictions of multi-task models against single-task models. We observe that in many cases multi-task learning performs better (correct or closer to the gold labels) than the single-task learning. A detailed analysis is depicted in Table XIV.

V. Conclusion and Future Work

In this work, we have proposed a multi-task ensemble framework for emotion analysis, sentiment analysis and intensity prediction. For ensemble we employed a MLP network that jointly learns multiple related tasks. First, we have developed three individual deep learning models (i.e., CNN, LSTM and GRU) to extract the learned representations. The multi-task ensemble network was further assisted through a hand-crafted feature vector. We evaluate our proposed approach on four benchmark datasets related to sentiment, emotion and intensity. Experimental results show that the multi-task framework is comparatively better than the single-task framework. Emotion detection can also be projected as multi-labeling task. However, due to absence of multi-emotion dataset we do not evaluate the proposed method on multi-emotion task.

In future we would like to extend our model for multi-label emotion classification. We would also like to evaluate the proposed model for the other related tasks.

VI. Acknowledgement

Asif Ekbal gratefully acknowledges JSPS Invitation Fellowship, Govt. of Japan, and Young Faculty Research Fel-

	Text	**Actual**	**Multi-task**	**Single-task**
EmoInt-2017	**Emotion Classification**			
	India should now react to the uri attack.	Anger	Anger	Fear
	what to wear Friday, speaking in front of 100's.	Fear	Fear	Joy
	Today you visited me in my dreams and even though you aren't physically gone I still mourn you.	Sad	Sad	Fear
	@user can't wait to see you Hun #cuddles #gossip.	Joy	Joy	Sad
	Intensity Prediction			
	Honestly don't know why I'm so unhappy most of the time. I just want it all to stop :(#itnevergoes.	Sad/0.94	Sad/0.59	0.49
	Everyday I wake up, a different @user player signs a contract extension! Love it!! #future #is #COYS	Joy/0.88	Joy/0.56	0.49
	@user that's a good question. I really don't know. I am slowly losing my optimism.	Joy/0.05	Joy/0.40	0.54
EmoBank	**Valence Prediction**			
	She bit the inside of her bottom lip, looking at me like I was breaking her poor, sweet heart.	2.0	2.91	3.54
	You've got it all wrong.	2.0	2.93	3.55
	Arousal Prediction			
	She caught herself, slid her jaw infinitesimally back into place, and said, "You don't like it?"	2.8	3.61	3.95
	He sighed and went on in softer, sadder voice, I guess I'm a thief now.	2.33	3.20	3.52
	Dominance Prediction			
	The region is poor (though not so much as I thought) and poorly serviced by internet and wifi: there's a router up the hill from her which sometimes provides wifi, but only then into a loft too unbearably hot to occupy during the day.	2.33	3.28	3.51
	He began to cry.	2.0	3.19	3.40
Facebook Posts	**Valence Prediction**			
	this is your life and its ending one minute at a time.	2.0	4.40	5.41
	I wonder why we go to school if education is knowledge, knowledge is power, power corrupts, corruption leads to crime and crime doesn't pay?	4.0	1.69	7.38
	I'm selfish	4.0	1.82	6.86
	I offically hate my brother.........Spartacus Blood and Sand. Incrediable	4.5	2.35	8.90
	Damn i hate this computers security system! its sooo paranoid it wont let me play a computer game since it is trying to access files saved on the computer!!!! =(	3.5	1.41	8.90
	Arousal Prediction			
	UGLY uniforms!!!!!!!!!!!!!!!! We look as bad as we play.	8.5	4.04	1.67
	yiiiiiiiiiiiippeeeeeeeeeeeeeeeee!!!!!!!	9.0	4.14	3.80
	Happy Thanksgiving	1.5	5.50	7.22

TABLE XIV: Error Analysis of Multi-task learning v/s Single-task learning.

lowship (YFRF), supported by Visvesvaraya PhD scheme for Electronics and IT, Ministry of Electronics and Information Technology (MeitY), Government of India, being implemented by Digital India Corporation (formerly Media Lab Asia).

References

[1] R. W. Picard, *Affective Computing*. Cambridge, MA, USA: MIT Press, 1997.

[2] P. Ekman, "An Argument for Basic Emotions," *Cognition and Emotion*, pp. 169–200, 1992.

[3] B. Pang, , and L. Lee, "Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with respect to Rating Scales," in *Proceedings of ACL*, 2005, pp. 115–124.

[4] M. D. Munezero, C. S. Montero, E. Sutinen, and J. Pajunen, "Are They Different? Affect, Feeling, Emotion, Sentiment, and Opinion Detection in Text," *IEEE transactions on affective computing*, vol. 5, no. 2, pp. 101–111, 2014.

[5] R. J. Davidson, K. R. Sherer, and H. H. Goldsmith, *Handbook of Affective Sciences*. Oxford University Press, 2009.

[6] J. A. Russell and L. F. Barrett, "Core Affect, Prototypical Emotional Episodes, and Other Things called Emotion: Dissecting the Elephant," *Journal of personality and social psychology*, vol. 76, no. 5, p. 805, 1999.

[7] C. Hawn, "Take Two Aspirin and Tweet Me in the Morning: How Twitter, Facebook, and other Social Media are Reshaping Health Care," *Health affairs*, vol. 28, no. 2, pp. 361–368, 2009.

[8] J. Bollen, H. Mao, and X. Zeng, "Twitter Mood Predicts the Stock Market," *Journal of computational science*, vol. 2, no. 1, pp. 1–8, 2011.

[9] G. Neubig, Y. Matsubayashi, M. Hagiwara, and K. Murakami, "Safety Information Mining - What can NLP do in a disaster -," in *Proceedings of 5th International Joint Conference on Natural Language Processing*, 2011, pp. 965–973.

[10] Z. Aldeneh, S. Khorram, D. Dimitriadis, and E. M. Provost, "Pooling acoustic and lexical features for the prediction of valence," in *Proceedings of the 19th ACM International Conference on Multimodal*

Interaction (ICMI-2017), 2017, pp. 68–72. [Online]. Available: http://doi.acm.org/10.1145/3136755.3136760

[11] S. Buechel and U. Hahn, "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*. Valencia, Spain: Association for Computational Linguistics, April 2017, pp. 578–585.

[12] D. Preoţiuc-Pietro, H. A. Schwartz, G. Park, J. Eichstaedt, M. Kern, L. Ungar, and E. Shulman, "Modelling Valence and Arousal in Facebook posts," in *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. San Diego, California: Association for Computational Linguistics, June 2016, pp. 9–15.

[13] S. Mohammad and F. Bravo-Marquez, "WASSA-2017 Shared Task on Emotion Intensity," in *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Copenhagen, Denmark: ACL, September 2017, pp. 34–49.

[14] B. Zhang, S. Khorram, and E. M. Provost, "Exploiting acoustic and lexical properties of phonemes to recognize valence from speech," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 5871–5875.

[15] R. Caruana, "Multitask Learning," *Machine Learning*, vol. 28, no. 1, pp. 41–75, Jul 1997.

[16] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, Oct. 2014, pp. 1746–1751.

[17] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[18] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling," *CoRR*, vol. abs/1412.3555, 2014.

[19] R. Collobert and J. Weston, "A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning," in *Proceedings of the 25th International Conference on Machine Learning*, ser. ICML '08. New York, NY, USA: ACM, 2008, pp. 160–167. [Online]. Available: http://doi.acm.org/10.1145/1390156.1390177

[20] A. Søgaard and Y. Goldberg, "Deep Multi-task Learning with Low Level Tasks Supervised at Lower Layers," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Berlin, Germany: Association for Computational Linguistics, August 2016, pp. 231–235. [Online]. Available: http://anthology.aclweb.org/P16-2038

[21] G. Balikas, S. Moura, and M.-R. Amini, "Multitask Learning for Fine-Grained Twitter Sentiment Analysis," in *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '17. New York, NY, USA: ACM, 2017, pp. 1005–1008.

[22] R. Xia and Y. Liu, "A Multi-Task Learning Framework for Emotion Recognition Using 2D Continuous Space," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 3–14, Jan 2017.

[23] G. Lu, X. Zhao, J. Yin, W. Yang, and B. Li, "Multi-task Learning using Variational Auto-Encoder for Sentiment Classification," *Pattern Recognition Letters*, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167865518302769

[24] A. Fraisse and P. Paroubek, "Toward a unifying model for Opinion, Sentiment and Emotion information extraction," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), May 2014, pp. 3881–3886. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2014/pdf/1010_Paper.pdf

[25] P. Ekman, *Basic Emotions*. The handbook of cognition and emotion., 1999.

[26] D. T. Ho and T. H. Cao, "A High-order Hidden Markov Model for Emotion Detection from Textual Data," in *Proceedings of the 12th Pacific Rim Conference on Knowledge Management and Acquisition for Intelligent Systems*, ser. PKAW'12, 2012, pp. 94–105.

[27] C.-H. Wu, Z.-J. Chuang, and Y.-C. Lin, "Emotion Recognition from Text Using Semantic Labels and Separable Mixture Models," *ACM ACM Transactions on Asian Language Information Pr ocessing*, vol. 5, no. 2, pp. 165–183, June 2006.

[28] P. Jain, P. Goel, D. Kulshreshtha, and K. K. Shukla, "Prayas at EmoInt 2017: An Ensemble of Deep Neural Architectures for Emotion Intensity Prediction in Tweets," in *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Copenhagen, Denmark: ACL, September 2017, pp. 58–65.

[29] M. Köper, E. Kim, and R. Klinger, "IMS at EmoInt-2017: Emotion Intensity Prediction with Affective Norms, Automatically Extended Resources and Deep Learning," in *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Copenhagen, Denmark: ACL, September 2017, pp. 50–57.

[30] M. S. Akhtar, P. Sawant, A. Ekbal, J. Pawar, and P. Bhattacharyya, "IITP at EmoInt-2017: Measuring Intensity of Emotions using Sentence Embeddings and Optimized Features," in *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Copenhagen, Denmark: Association for Computational Linguistics, September 2017, pp. 212–218.

[31] P. Xu, A. Madotto, C.-S. Wu, J. H. Park, and P. Fung, "Emo2Vec: Learning Generalized Emotion Representation by Multi-task Training," in *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Brussels, Belgium: Association for Computational Linguistics, October 2018, pp. 292–298.

[32] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global Vectors for Word Representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543.

[33] D. Tang, F. Wei, B. Qin, N. Yang, T. Liu, and M. Zhou, "Sentiment Embeddings with Applications to Sentiment Analysis," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 2, pp. 496–509, Feb 2016.

[34] B. Felbo, A. Mislove, A. Søgaard, I. Rahwan, and S. Lehmann, "Using Millions of Emoji Occurrences to Learn Any-Domain Representations for Detecting Sentiment, Emotion and Sarcasm," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, September 2017, pp. 1615–1625. [Online]. Available: https://www.aclweb.org/anthology/D17-1169

[35] M. Kabadjov, A. Balahur, and E. Boldrini, "Sentiment Intensity: Is It a Good Summary Indicator?" in *Human Language Technology. Challenges for Computer Science and Linguistics*, Z. Vetulani, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 203–212.

[36] S. Kiritchenko, S. Mohammad, and M. Salameh, "SemEval-2016 Task 7: Determining Sentiment Intensity of English and Arabic Phrases," in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. San Diego, California: Association for Computational Linguistics, June 2016, pp. 42–51. [Online]. Available: http://www.aclweb.org/anthology/S16-1004

[37] R. Sharma, A. Somani, L. Kumar, and P. Bhattacharyya, "Sentiment Intensity Ranking among Adjectives Using Sentiment Bearing Word Embeddings," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, September 2017, pp. 547–552.

[38] M. S. Akhtar, A. Kumar, D. Ghosal, A. Ekbal, and P. Bhattacharyya, "A Multilayer Perceptron based Ensemble Technique for Fine-grained Financial Sentiment Analysis," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, September 2017, pp. 540–546.

[39] Y. Freund and R. E. Schapire, "Experiments with a New Boosting Algorithm," in *Thirteenth International Conference on Machine Learning*. San Francisco: Morgan Kaufmann, 1996, pp. 148–156.

[40] L. Breiman, "Bagging Predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[41] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On Combining Classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 3, pp. 226–239, Mar. 1998. [Online]. Available: http://dx.doi.org/10.1109/34.667881

[42] A. Ekbal and S. Saha, "Weighted Vote-Based Classifier Ensemble for Named Entity Recognition: A Genetic Algorithm-Based Approach," *ACM Transactions on Asian Language Information Processing*, vol. 10, no. 2, pp. 9:1–9:37, June 2011.

[43] T. Xiao, J. Zhu, and T. Liu, "Bagging and Boosting Statistical Machine Translation Systems," *Artif. Intell.*, vol. 195, pp. 496–527, Feb. 2013.

[44] K. R. Remya and J. S. Ramya, "Using Weighted Majority Voting Classifier Combination for Relation Classification in Biomedical Texts," in *2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT)*, July 2014, pp. 1205–1209.

[45] M. S. Akhtar, D. Gupta, A. Ekbal, and P. Bhattacharyya, "Feature Selection and Ensemble Construction: A Two-Step Method for Aspect Based Sentiment Analysis," *Knowledge-Based Systems*, vol. 125, pp. 116 – 135, 2017.

[46] J. Gideon, S. Khorram, Z. Aldeneh, D. Dimitriadis, and E. M. Provost, “Progressive Neural Networks for Transfer Learning in Emotion Recognition,” *CoRR*, vol. abs/1706.03256, 2017.

[47] T. G. Dietterich, “Ensemble Methods in Machine Learning,” in *Proceedings of the First International Workshop on Multiple Classifier Systems*, ser. MCS ’00. London, UK, UK: Springer-Verlag, 2000, pp. 1–15. [Online]. Available: http://dl.acm.org/citation.cfm?id=648054.743935

[48] X. Glorot, A. Bordes, and Y. Bengio, “Deep Sparse Rectifier Neural Networks,” in *Aistats*, vol. 15, no. 106, 2011, p. 275.

[49] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, “Improving neural networks by preventing co-adaptation of feature detectors,” *CoRR*, vol. abs/1207.0580, 2012.

[50] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” *CoRR*, vol. abs/1412.6980, 2014.

[51] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A Simple Way to Prevent Neural Networks from Overfitting,” *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[52] J. Wiebe and R. Mihalcea, “Word Sense and Subjectivity,” in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2006, pp. 1065–1072.

[53] X. Ding, B. Liu, and P. S. Yu, “A Holistic Lexicon-based Approach to Opinion Mining,” in *Proceedings of the 2008 international conference on web search and data mining*. ACM, 2008, pp. 231–240.

[54] S. Mohammad, S. Kiritchenko, and X. Zhu, “NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets,” in *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. Atlanta, Georgia, USA: Association for Computational Linguistics, June 2013, pp. 321–327.

[55] F. Å. Nielsen, “A new ANEW: Evaluation of a Word List for Sentiment Analysis in Microblogs,” *CoRR*, vol. abs/1103.2903, 2011.

[56] S. Baccianella, A. Esuli, and F. Sebastiani, “SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining,” in *LREC*, vol. 10, 2010, pp. 2200–2204.

[57] S. M. Mohammad and P. D. Turney, “Crowdsourcing a Word-Emotion Association Lexicon,” *Computational Intelligence*, vol. 29, no. 3, pp. 436–465, 2013.

[58] F. Bravo-Marquez, E. Frank, S. M. Mohammad, and B. Pfahringer, “Determining Word–Emotion Associations from Tweets by Multi-label Classification,” in *WI’16*. IEEE Computer Society, 2016, pp. 536–539.

[59] S. M. Mohammad and S. Kiritchenko, “Using Hashtags to Capture Fine Emotion Categories from Tweets,” *Computational Intelligence*, vol. 31, no. 2, pp. 301–326, 2015.

[60] C. H. E. Gilbert, “VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text,” in *Proceedings of the Eighth International Conference on Weblogs and Social Media (ICWSM-14).*, 2014, pp. 216–225.

[61] P. Nakov, A. Ritter, S. Rosenthal, F. Sebastiani, and V. Stoyanov, “SemEval-2016 task 4: Sentiment analysis in twitter,” in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, San Diego, California, June 2016, pp. 1–18.

[62] F. Chollet *et al.*, “Keras,” https://keras.io, 2015.

[63] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine Learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[64] L. A. M. Bostan and R. Klinger, “An Analysis of Annotated Corpora for Emotion Classification in Text,” in *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA, August 2018, pp. 2104–2119.

[65] J. H. Friedman, “Greedy function approximation: a gradient boosting machine,” *Annals of statistics*, pp. 1189–1232, 2001.

[66] Y. Freund and R. E. Schapire, “A desicion-theoretic generalization of on-line learning and an application to boosting,” in *European conference on computational learning theory*. Springer, 1995, pp. 23–37.

[67] L. Breiman, “Bagging predictors,” *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.

Md Shad Akhtar is a research scholar in the Department of Computer Science and Engineering, IIT Patna. He completed his M.Tech form IIT(ISM), Dhanbad in 2014 and BEng from JMI, New Delhi in 2009. His main area of research is Natural Language Processing and Sentiment Analysis. He has published papers in various peer reviewed conferences and journals of international repute.

Deepanway Ghosal is an undergraduate student in the Department of Electrical Engineering, IIT Patna. His areas of research are Natural Language Processing and Multi-modal Machine Learning.

Dr. Asif Ekbal is currently an Associate Professor in the Department of Computer Science and Engineering, IIT Patna. He has been pursuing research in Natural Language Processing, Information Extraction, Text Mining and Machine Learning applications for the last 11 years. He is involved with different sponsored research projects in the broad areas of Artificial Intelligence and Machine Learning technologies, funded by different Govt. and private agencies.

Prof. Pushpak Bhattacharyya is the Director of IIT Patna and a Professor of Computer Science and Engineering, IIT Patna and IIT Bombay. He is an outstanding researcher in Natural Language Processing and Machine Learning. He has contributed to all areas of Natural Language Processing, encompassing machine translation, sentiment and opinion mining, cross lingual search and multilingual information extraction. He had held the office of Association for Computational Linguistics (ACL), the highest body of NLP, as its president. He received several awards and fellowships such as Fellow of Indian National Academy of Engineering, IBM Faculty Award, Yahoo Faculty Award etc.

Prof. Sadao Kurohashi received the B.S., M.S., and PhD in Electrical Engineering from Kyoto University in the years 1989, 1991 and 1994, respectively. He has been a visiting researcher of IRCS, University of Pennsylvania in 1994. He is currently a professor of the Graduate School of Informatics at Kyoto University. He has presented scientific papers in many national and international conferences and published in various journals and books in the field of Natural Language Processing. His current interests include machine translation, information retrieval (a principal member of New IT Infrastructure for the Information-explosion Era by MEXT), and knowledge engineering.