A Computational Model to Understand Emotions in Sarcasm

Anonymous EMNLP submission

Abstract

Sarcasm is generally associated with a negative emotion. The question is *which negative emotion- anger, sadness, disgust, any other?* This paper presents a methodology of detecting the exact emotion(s) in a sarcastic sentence. Sarcasm arises from contextual incongruity in a sentence and bears a surface sentiment which is different from the intended sentiment. While the surface sentiment may be positive, the intended sentiment is negative. Thus the underlying emotion recognition task becomes one of the most difficult parts of the conundrum. Previous works have extensively studied sentiment and emotion in language, while the relationship between sarcasm and emotion has been largely unaddressed. In order to take a principled approach towards studying this relationship, we introduce to the community the first benchmark dataset of annotated sarcasm in videos with 8 primary emotions, arousal and valence levels. Leveraging Plutchik's wheel of emotions and arousal prediction, we infer 32 emotions, without explicitly needing to label all the data manually. Specifically, we predict 24 emotions using the 8 predicted primary emotions and arousal levels. Further, we infer 8 high-level combination emotions that arise from the presence of multiple primary emotions. Our baseline results show that by utilizing the sarcasm label as an input, hamming loss is decreased by 8%, and the micro f-score increases significantly for most emotions. To the best of our knowledge, this is the first work on sarcasm in emotion recognition and first such dataset for use by the research community.

1 Introduction

Sarcasm is a very sophisticated linguistic articulation where the sentential meaning is often disbelieved due to the linguistic incongruencies or differences in implied and surface sentiment. While incongruity is the key element of sarcasm, the intent could be to appear humorous, ridicule someone, or express contempt. Thus sarcasm is often considered a very nuanced, creative, or intelligent language construct which poses several challenges to both detection and generation.

Detecting emotions and sarcasm is crucial for all services involving human interactions, such as chatbots, e-commerce, e-tourism and several other businesses. We hypothesize that sarcasm affects the emotion associated with a conversation and thus this paper aims to study the emotions, arousal and valence in sarcastic sentences.Valence measures the positive or negative affectivity. Arousal measures the intensity of the emotion associated (Cowie and Cornelius, 2003). To the best of our knowledge, there exist no works that have studied sarcasm and emotion together, thus no such datasets are available. Towards this direction, we manually prepare a benchmark dataset 'emo-UStARD' with 8 emotions and arousal-valence labels. This dataset is an extension of MUStARD data (Castro et al., 2019) which is a Multimodal Sarcasm Detection dataset of 690 video instances with contextual videos of the dialogue collected from English TV series.

We perform exhaustive experimentation for training multi-label emotion classifiers: (a) using only utterances, (b) using utterance with contextual information of the previous spoken dialogues, (c) using utterance with sarcasm/non-sarcasm label and (d) using all the inputs together. We used pretrained BERT (Devlin et al., 2018) word embeddings and used transfer learning (Bengio, 2012) to train models on other larger datasets for emotion recognition such as CMU-MOSEI (Zadeh et al., 2018) and IEMOCAP (Busso et al., 2008) for creating baseline models. Since these datasets do not have sarcasm labels or contextual sentences, we could utilize them in experiments using only utterance. We fine tuned the pretrained model on

our proposed dataset emo-UStARD for the other experiments using sarcasm label and context as inputs.We build baselines classifiers for multi-label and single label setting.

Along with the 8 primary emotions, we also infer 24 other emotions using the rules from Plutchik's wheel of emotion (Plutchik, 1991). The emotions which vary from the primary emotion only in terms of intensity are referred to intensified emotions in the rest of the paper. For example, pensiveness and grief are lower and upper intensity levels of sadness as emotions. There are 16 such intensified emotions which we infer using predictions of arousal and primary emotions. The emotions which arise from the combination of two primary emotions are being referred to as combination emotion. We write rules using Plutchik's wheel to identify the eight combination emotions. We present the frequency distribution of emotions (primary, high intensity and combination), present in sarcastic versus non-sarcastic sentences of our dataset.

Contributions of this paper are:

- A benchmark dataset 'emo-UStARD', of sarcastic and non-sarcastic videos, that is annotated with 8 primary emotions, and also arousal and valence levels to get the intensity of emotions.
- Emotion recognition classifiers trained using utterances, dialogue contextual sentences, and sarcasm label to study the influence of sarcasm on emotion recognition.
- We use arousal predictions along with primary emotion recognition to infer 24 intensity varying emotions, and eight combination emotion using rules from Plutchik's 3D wheel of emotions.

2 Related work

Research studying the impact of sarcasm on sentiment analysis (Maynard and Greenwood, 2014) showed that sarcasm often has a negative sentiment, but the associated emotions have not been studied. For tweet analysis, NLP researchers have tried to detect sarcasm and perform sentiment analysis together (Poria et al., 2016; Bharti et al., 2015), while some try to improve sentiment analysis performance using sarcasm detection (Bouazizi and Ohtsuki, 2015).

Figure 1: Plutchik's wheel of Emotions: 3D representation showing 8 primary emotions in the center circle with their intensity variants and combinations forming 32 emotions.

There exists extensive cross-disciplinary research on human emotions that have established several scales of emotions. Ekman's scale (Ekman, 1999) of six basic emotions is one such model and several data sets have used this scale for emotion recognition (Zadeh et al., 2018, 2016). Ekman's six basic emotions are: *Joy, Sadness, Surprise, Anger, Disgust* and *Fear*. Plutchik's wheel of emotions (Plutchik, 1991) is another popular wheel structure that represents 32 emotions, with eight primary emotions, three intensity variants of each primary emotion and eight combination emotions (combination of two primary emotions shown between the petals) as shown in Figure 1. Basic emotions according to Plutchik's wheel are *joy*, *trust*, *fear*, *surprise*, *sadness*, *disgust*, *anger*, and *anticipation*, wherein trust and anticipation were the two new emotions introduced over and above Eckman's. Opposite emotions are placed diametrically opposite to each other, such as joy and sadness, or anticipation and surprise. Each 'petal' of the wheel indicates the arousal of the emotion, with darker colour indicating higher intensity.

We annotated the dataset with the 8 basic emotions from the Plutchik wheel of emotions. This helped us in build computational models to infer the intensified emotions and the highly nuanced combination emotions.

3 Proposed Dataset: emo-UStARD

While there exist a few data sets for sarcasm detection (Riloff et al., 2013; Ptáček et al., 2014), sarcasm and emotion have not been studied together before. MUStARD (Castro et al., 2019) is the first multimodal data set annotated for sarcasm detec-

tion task. This data contains balanced set of 345 sarcastic, and 345 non-sarcastic video utterances. Each utterance has one or two contextual video utterances for better understanding. This dataset contains a subset of Multimodal Emotion Lines Dataset (MELD) data set (Poria et al., 2018) which is a multimodal extension of EmotionLines data set(Chen et al., 2018). MELD contains about 13,000 utterances from the TV-series Friends, labeled with one of the seven emotions (anger, disgust, sadness, joy, neutral, surprise and fear) and sentiment. EmotionLines (Chen et al., 2018) is a textual data set comprising of 29,245 utterances from the series Friends and private Facebook messenger dialogues. In this study, we first propose a benchmark data set emo-UStARD built using the MUStARD videos (Castro et al., 2019).

3.1 Annotation Process

The proposed data set is annotated manually by 5 annotators out of whom we had one professional linguist and 4 graduate students working in the area of emotion and sentiment recognition. Thus all our annotators were very familiar with the task. Each annotator could give multiple emotion labels to the utterances, along with arousal, valence and their confidence in annotating the particular utterance. Since contextual information plays a very important role in determining the associated emotions (Busso et al., 2008; Cauldwell, 2000), the contextual videos were observed by the annotators while annotating the utterance in the video. The data set has multiple emotions associated with each utterance like most natural human interactions. Thus we assign multiple labels of emotion rather than choosing one single emotion.

Previous emotion recognition works (Poria et al., 2018; Chen et al., 2018) have considered assigning one value of emotion for ease of recognition. To compare with CMU-MOSEI and Iemocap (Tripathi et al., 2018), we build models using single emotion label. These models are tested on emo-UStARD as well, by considering top common emotion from all annotators as the single emotion label. For CMU-MOSEI (Zadeh et al., 2018), there were 3 annotators who gave their confidence values while labeling. We discarded low confidence utterances and picked utterances with confidence score of annotation to be greater than 1. This resulted in a data subset of 6245 instances for training (full training set was 16327), 599 instances for validating (out of 1871) and 1755 for testing (out of 4462).

3.2 Inter Annotator Agreement

The confidence values of annotators help us in resolving conflicts and choosing the grountruth labels. The confidence values are also an indicator of the subjectivity and challenges of emotion recognition in presence of sarcasm. We used Krippendorff's alpha algorithm (Krippendorff, 2011) which is suitable for multi-label inter-annotator reliability. For the 8 primary emotion labels, our average of 5 annotators is 0.46. Such an average value of krippendorf's alpha, indicates that sarcasm makes annotation very challenging and the perceived emotions can be very subjective.

For inter-annotator agreement on arousal and valence, we used Cronbach alpha coefficient (Cronbach, 1951) as used by other data sets such as Iemocap. The average Cronbach alpha coefficient for arousal and valence annotation is 0.39 and 0.43. The trend here is similar to Iemocap dataset (Busso et al., 2008) where authors reported a higher agreement in valence annotation than arousal levels.

3.3 Dataset Statistics

Detailed data set statistics is given in Table 1.

Table 1: Dataset Statistics

Total number of utterances	690
Average length of utterance	14 tokens
Average duration of utterance	5.22 seconds
Maximum length of utterance	73 tokens
Total number of unique words in dataset	1991
Total number of emotion labels	8
Utterances with single emotion present	334
Utterances with 2 emotion labels	296
Utterances with 3 or more emotion labels	60

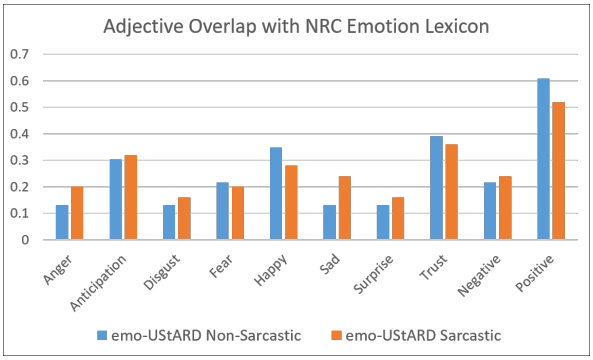

Figure 2: Adjective Overlap of emo-UStARD dataset with EmoLex [1]

To get more insights into the dataset, we compute word overlap and adjective overlap with popu-

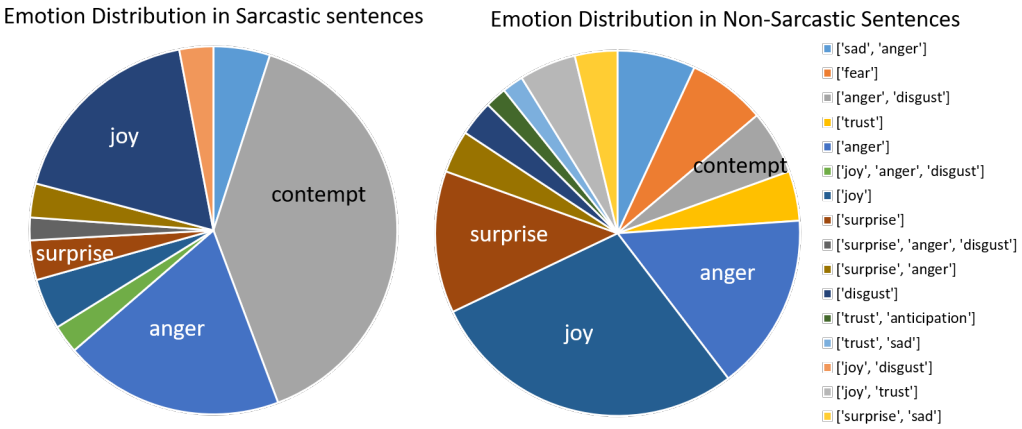

Figure 3: Top Frequency Emotions present in Sarcastic versus Non-sarcastic sentences.

lar NRC Emotion Lexicon EMOLEX(Mohammad and Turney, 2010). Adjectives are strong markers of emotion, thus we show the distribution of adjectives for each emotion as well as positive and negative sentiment on the emo-UStARD dataset in Figure 2. As seen in Figure 2, sarcastic sentences have more occurrence of anger, sadness, disgust, anticipation and surprise than joy, trust or fear. Since sarcasm is characterized by positive *surface* sentiment, and negative *intended* sentiment, word match with lexicons show that the positivity in sarcastic sentences is quite high. We also observe that the negative sentiment in sarcastic sentences is higher than the non-sarcastic sentences.

While analyzing the human annotations, we observe a set of 61 unique label combinations have been assigned to the utterances. Since there are 690 utterances, 61 label combinations lead to a long tail distribution. To identify the labels with higher frequency, we filter out emotions that have appeared once and twice, thus getting top 31 label combinations. Figure 3 shows the distribution of these top 31 label combinations in sarcastic (left) and non-sarcastic sentences (right). We see the class ['anger','disgust'] with largest number of utterances among sarcastic sentences. This is a combination emotion: 'contempt' according to Plutchik's wheel, which is a negative sentiment. While 'anger' and 'joy' can be seen in both sarcastic and non-sarcastic sentences, most sarcastic sentences have negative emotions.

We wanted to study the arousal levels for sarcastic and non-sarcastic sentences. Arousal and valence both have been labeled in the range of -1 to 1. The average of arousal labels is 0.31 for sarcastic sentences and 0.22 for non-sarcastic sentences. This indicates that higher level of excitement triggers sarcasm. The average of valence labels is -0.18 for sarcasm and 0.01 for non-sarcastic utterances, indicating negative affectivity in sarcastic utterances. As seen in Figure 3, for emo-UStARD, there is a huge proportion of sentences which have anger, or disgust as the associated emotion for both sarcastic and non-sarcastic sentences. Thus the averages of arousal and valence for sarcastic and non-sarcastic sentences are in close range. This distribution of sentences affects prediction of arousal, valence and increases classifier confusion while recognizing emotions.

4 Experimental Setup

We designed four experiments to study emotion recognition in sarcastic or non-sarcastic sentences. Since we have multiple emotions for each utterance, we perform experiments in both multi-label classification setting and one-vs-rest setting (Binary Relevance method). Initially we used emo-UStARD data to build a classifier for all 8 primary emotions using BERT word embeddings and a multi-layer perceptron. Since the data set is small, the recall was low and often zero for some emotions, thus compelling us to use other emotion data sets to bootstrap learning.

The **first experiment** uses utterance as the only input. For this experiment, we used popular emotion data sets such as CMU-MOSEI and Iemocap for training the model. We tested the model directly on their test sets as well as on emo-UStARD's sarcasm and non-sarcasm subsets. The **second experiment** uses the sarcasm label along with the BERT word embeddings (Devlin et al., 2018) of the sentences. The **third experiment** uses the utterance

with the contextual utterances of the dialogue. In the **fourth experiment** we use the utterance along with the sarcasm label, and contextual sentences for emotion prediction. Since sarcasm and contextual information are not present in CMU-MOSEI and Iemocap datasets, we fine tuned the learnt model of first experiment on the emo-UStARD train subset for these three experiments. We also performed the **same four experiments** for arousal and valence predictions. For all experiments, we use BERT tokenizer and pass 768 dimensional BERT word embeddings to an LSTM based classifier, which uses Adam optimizer (Kingma and Ba, 2014), with a dropout of 0.5, and early stopping criterion for regularization.

Evaluation Metrics

We use standard multi-label classification metrics such as *micro-precision, micro-recall, micro-f-score, hamming loss* and *subset accuracy*. Micro-averaging based measures help understand the performance of a system across sets of data and is especially useful when the data varies in size, as in our case. Hamming loss gives the fraction of labels that were incorrectly predicted. Subset accuracy being a strict metric, provides lower bound of the system showing the percentage of samples that have all labels classified correctly.

5 Results and Analysis

Exhaustive evaluation of emotion recognition is performed in different settings to create baselines for the benchmark database 'emo-UStARD'. We also build regression models for arousal and valence and use the arousal predictions to infer the other emotions.

Table 2 shows the mean square error (MSE) and mean absolute error (MAE) in the prediction of arousal and valence using utterance, utterance and sarcasm, utterance and context as well as utterance, sarcasm and context as the four different inputs. Error values remain almost unaffected with the addition of sarcasm label as an input. However, the error decreases when the context information is utilized, for both arousal and valence prediction. While fine tuning we pass the BERT word embeddings to a 4 layer LSTM network using 7:1:2 partitions of train, valid and test of emo-UStARD. The final numbers reported use a learning rate of 0.0005, for 150 epochs, while choosing the model at the lowest MSE.

Table 2: Test results for Arousal-Valence prediction on 4 experiments on emo-UStARD testset using different inputs; model trained on select subset of CMU-MOSEI dataset and IEMOCAP and fine tuned with the emo-UStARD train subset. Metrics used: MSE (mean square error) and MAE (Mean absolute error)

Inputs	Metrics	Arousal	Valence
utterance	MSE	0.16	0.29
	MAE	0.31	0.44
: utterance, sarvec	MSE	0.15	0.28
:	MAE	0.32	0.43
utterance, context	MSE	0.13	0.24
:	MAE	0.28	0.41
utterance, sarvec, context	MSE	0.12	0.28
:	MAE	0.27	0.43

Both CMU-MOSEI and Iemocap have single emotion per sentence. While CMU-Mosei has 6 emotions: Joy, Sad, Anger, Surprise,Fear and Disgust, IEMOCAP has seven, anticipation being the additional emotion. Table 3 shows model trained on subsets of CMU-MOSEI and IEMOCAP, tested on emo-UStARD.

For experiments using only utterance, we first trained an LSTM classifier on the BERT word embeddings of Mosei and Iemocap data. We test this baseline model on Mosei and Iemocap test set as well as the full multi-label emo-UStARD data (sarcastic and non-sarcastic partitions of 345 videos each) as shown in Table 3. The model performs reasonably well on cmu-mosei and Iemocap testset with the exact match (subset accuracy) of 56.58% and partial match of 90% (computed from hamming loss).

For emo-UStARD, the same model has an exact match of 10.72%. One reason for this significant drop in subset accuracy (exact match) is the size of emo-UStARD utterances which are much smaller in comparison to full sentences in Mosei or Iemocap. Hamming loss measures the fraction of labels predicted incorrectly and is very appropriate for multilabel classification setting like emo-UStARD. Hamming loss is in the range of 0-1, where 0 means exact match and 1 means no labels matched. For sarcastic sentences, although the hamming loss is 0.29, the subset accuracy (exact match) is only 7.82%, manifesting the challenges posed by sarcasm. Incase of utterances with multiple emotions,

Table 3: Test results for Experiment 1: Emotion recognition using only utterances as input. Model is trained on high confidence emotion labels of CMU-MOSEI dataset and IEMOCAP. Multi-label micro-averaged metrics are shown under each emotion. The column 'Overall' contains subset-accuracy and hamming loss computed for multi-label classification. The metric Accuracy reports accuracy for single label emotion prediction on same test subsets.

Test	Metrics	Joy	Sad	Anger	Surprise	Disgust	Fear	Anti	Overall
Testsubset.	Precision	0.67	0.68	0.64	0.70	0.45	0.84	0.80	-
	Recall	0.83	0.56	0.39	0.61	0.32	0.67	0.55	-
	F-score	0.74	0.62	0.48	0.65	0.31	0.75	0.65	-
	Subset Accuracy	-	-	-	-	-	-	-	56.58%
	Hamming Loss	-	-	-	-	-	-	-	0.0996
	Accuracy	-	-	-	-	-	-	-	**56.58%**
emo-UStARD NonSar.	Precision	0.43	0.32	0.31	0	0	0.33	0.20	-
	Recall	0.03	0.23	0.30	0	0	0.03	0.30	-
	F-score	0.05	0.26	0.31	0	0	0.05	0.24	-
	Subset Accuracy	-	-	-	-	-	-	-	**10.72%**
	Hamming Loss	-	-	-	-	-	-	-	**0.2525**
	Accuracy	-	-	-	-	-	-	-	**23.76%**
emo-UStARD Sar.	Precision	0.19	0.11	0.72	0	0	0	0.04	-
	Recall	0.08	0.21	0.27	0	0	0	0.36	-
	F-score	0.12	0.14	0.39	0	0	0	0.08	-
	Subset Accuracy	-	-	-	-	-	-	-	**7.82%**
	Hamming Loss	-	-	-	-	-	-	-	**0.2890**
	Accuracy	-	-	-	-	-	-	-	**17.40%**

we observe that one out of two emotions is always predicted correctly. For utterances with three or more emotions, 70% utterances have two out of three emotions correctly predicted, especially when the emotions are conflicting emotions such as joy and disgust, anger and surprise etc.

Table 4 shows all the four experiments where a model is first pretrained using utterances from CMU-MOSEI and Iemocap and then finetuned on emo-UStARD train partition using different inputs for different experiments. We perform 5-fold cross validation to prevent overfitting, since the data set size is small and there is severe class imbalance as seen in 3. We use 4 layer network (2048,1024,512,128 cells) with ReLU activation, Adam optimizer (Kingma and Ba, 2014) and dropout of 0.5. The learning rate and number of epochs were selected by uniform sampling within the range of [0.0001 to 0.00009] and [100-300] for learning rate and number of epochs respectively for both pretraining and finetuning. As seen in first row of the table, when fine tuned with utterances from emo-UStARD dataset, the exact match accuracy for all emotions is 9.27%, while the hamming loss is 0.28. This indicates that for 70% utterances, predicted emotions partially match the groundtruth set of labels, but only for 10% all the labels present in grountruth exactly match. For the second experiment, when the information of the sentence being sarcastic or non-sarcastic is passed, we observe a slight increase in subset accuracy as well as the micro f-score for each emotion. When contextual sentences are passed as input, the overall metrics improve in comparison to only utterance based experiments. However, for this data, we observe that utterance and sarcasm label served as a better input than utilizing contextual sentences for emotion prediction. Our initial hypothesis that the sarcasm label and the context would help the classifier was based on the importance of context for even human annotation. Since the contextual sentences might have a complete different emotion than the test utterance, their word embeddings might be confusing the classifier.

Table 4: Test results for 4 experiments on emo-UStARD test set using different inputs; model trained on select subset of CMU-MOSEI dataset and IEMOCAP and finetuned with the emo-UStARD train subset. Note:The column Anti refers to Anticipation emotion

Inputs	Metrics	Joy	Sad	Anger	Surprise	Disgust	Fear	Anti	Overall
utterance :	Precision	0.26	0.24	0.60	0.41	0.40	0	0.12	-
	Recall	0.15	0.40	0.25	0.08	0.01	0	0.30	-
	F-score	0.19	0.30	0.35	0.13	0.01	0	0.17	-
	Subset Accuracy	-	-	-	-	-	-	-	9.27%
	Hamming Loss	-	-	-	-	-	-	-	0.28
utterance, sar-vec :	Precision	0.15	0.26	0.59	0.24	0	0	0.20	-
	Recall	0.08	0.39	0.18	0.13	0	0	0.27	-
	F-score	0.10	0.31	0.27	0.17	0	0	0.22	-
	Subset Accuracy	-	-	-	-	-	-	-	10.97%
	Hamming Loss	-	-	-	-	-	-	-	0.2
utterance, context :	Precision	0.32	0.35	0.61	0.20	0	0.1	0.15	-
	Recall	0.06	0.15	0.16	0.03	0	0.05	0.46	-
	F-score	0.09	0.21	0.25	0.06	0	0.07	0.23	-
	Subset Accuracy	-	-	-	-	-	-	-	10.82%
	Hamming Loss	-	-	-	-	-	-	-	0.26
utterance, sar-vec, context	Precision	0.38	0.41	0.59	0.14	0	0.10	0.14	-
	Recall	0.05	0.15	0.13	0.03	0	0.02	0.42	-
	F-score	0.09	0.22	0.24	0.05	0	0.03	0.20	-
	Subset Accuracy	-	-	-	-	-	-	-	10.74%
	Hamming Loss	-	-	-	-	-	-	-	0.24

Emo-UStARD has a high imbalance ratio, typical of most benchmark multilabel datasets. Thus we use Binary relevance method (OneVsRest classification in multilabel setting) which handles imbalances. Table 5 shows results of OnevsRest classifiers using different training data and features. For baseline, we experimented with term-frequency and inverse-document frequency based features (tf-idf) (Salton and Buckley, 1987) using linear support vector classifier. The first row uses class-balanced data from CMU-Mosei for train, thus has results for six emotions only. The second and third uses Mosei and Iemocap train sets but the third row uses a BERT embeddings instead of tf-idf features. We observe that linear SVC using balanced CMU-Mosei for training gave best results for both sarcastic and non-sarcastic sentences in the onevs-Rest setting. This can be attributed to the fact that train data of Mosei and Iemocap is already highly skewed with 50% of the data belonging to Joy, while fear and surprise together comprise of 6%, and disgust, and anticipation are 12% each. We compute subset accuracy as shown in Table 5 as well as microaveraged precision, recall, fscore presented in appendix.

Table 6 shows a few examples of high confidence annotations of sarcastic sentences with the predicted primary emotions, arousal values, intensified emotions, and combination emotions. Row 1 shows an utterance with very high arousal leading to inference of an intensified emotion of anger. Row 3 is an example of sarcasm made for mockery and shows the contrastive emotions, a typical property of sarcasm. The last row shows an utterance with primary emotions of sadness, anger and disgust with very low values of arousal. While the average arousal for sarcastic sentences is 0.31, only 14% of utterances have arousal below average (zero) indicating that sarcasm need not be always at an excited level of emotion. Also for this case, using the rules we infer two combination emotions: *remorse* arising from sadness and disgust; *contempt*

Table 5: Accuracy results on emo-UStARD full data (sarcasm (emo-UStARD Sar) and non-sarcasm partitions (emo-UStARD Non-sar)) using OneVsRestClassifiers. the column 'Anti' indicates anticipation emotion.

Classifier	Test Subset	Joy	Sad	Anger	Surprise	Disgust	Fear	Anti
Linear SVC .	CMU-test	49.68	84.38	78.97	93.90	84.44	90.25	-
	emo-UStARD Sar	74.49	82.02	**42.60**	76.23	**53.33**	90.72	-
	emo-UStARD Non-sar	**54.78**	76.52	82.60	72.46	91.88	86.95	-
Linear SVC .	mosei+Iemocap test	85.6	82.6	82	94.3	87.6	96.5	94
	emo-UStARD Sar	79.4	84.3	**36.5**	83.2	**35.4**	95.9	94.6
	emo-UStARD Non-sar	**63.2**	71.9	72.1	76.8	87.8	88.9	76.8
BERT LSTM Classifier	Mosei+iemocap test	79.3	18.4	70.3	93.7	18.4	8.9	79.6
	emo-UStARD Sar	85.5	**22.8**	40.5	83.7	65.8	**7.3**	94.8
	emo-UStARD Non-sar	64.6	23	65.2	75.9	11	12.8	79.1

Table 6: Examples from the dataset with primary emotions, arousal, higher order emotion and combination emotion.

Utterances	Primary Emotion	Arousal	Intensified Emotion	Combination Emotion
I am not freaking out, why would I be freaking out? A woman named Hildy called and said we will get married, but that happens everyday	['anger']	0.833	['loathing']	-
No, you're right, we should do what you do. Have our mom send us pants from the Walmart in Houston.	['anger', 'disgust']	0.667	-	contempt
Leonard's work is nearly as amazing as third graders growing lima beans in wet paper towels.	['joy', 'disgust']	0.233	-	-
Oh! Satan's minions at work again?	['surprise']	0.566	['amazement']	-
When I didn't pay my bill, the Department of Water and Power thought I would enjoy the ambience.	['sad', 'anger', 'disgust']	-0.533	-	['remorse','contempt']

arising from disgust and anger.

6 Summary

This paper presents a method for emotion understanding in sarcastic sentences. It also gives a benchmark data set to help the task. The said data set is the first to have sarcasm and emotions labeled together that should help the community. Sarcasm poses different challenges such as surface positivity and intended or perceived negativity, making emotions very subjective and difficult to recognize. We build a suite of baseline classifiers using different inputs, and features to study the influence of each input in emotion recognition.

7 Conclusions and Future Work

The proposed dataset would enable conducting different kinds of studies on sarcasm, emotion and sentiment. Using arousal predictions and rules from Plutchik's wheel structure,we are able to infer 32 emotions from labeled data of eight emotions, with several high level combination emotions that are considered holy grail for automated emotion recognition. Since some emotions are better perceived from audio or video, the authors aim to leverage those modalities present in the proposed data set to improve recognition rates of this task as a future work by using both verbal and non-verbal cues.

References

Yoshua Bengio. 2012. Deep learning of representations for unsupervised and transfer learning. In *Proceedings of ICML workshop on unsupervised and transfer learning*, pages 17–36.

Santosh Kumar Bharti, Korra Sathya Babu, and Sanjay Kumar Jena. 2015. Parsing-based sarcasm sentiment recognition in twitter data. In *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 1373–1380. IEEE.

Mondher Bouazizi and Tomoaki Ohtsuki. 2015. Opinion mining in twitter how to make use of sarcasm to enhance sentiment analysis. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, pages 1594–1597.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards multimodal sarcasm detection (an _obviously_ perfect paper). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Richard T Cauldwell. 2000. Where did the anger go? the role of context in interpreting emotion in speech. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*.

Sheng-Yeh Chen, Chao-Chun Hsu, Chuan-Chun Kuo, Ting-Hao K. Huang, and Lun-Wei Ku. 2018. Emotionlines: An emotion corpus of multi-party conversations. *CoRR*, abs/1802.08379.

Roddy Cowie and Randolph R Cornelius. 2003. Describing the emotional states that are expressed in speech. *Speech communication*, 40(1-2):5–32.

Lee J Cronbach. 1951. Coefficient alpha and the internal structure of tests. *psychometrika*, 16(3):297–334.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Paul Ekman. 1999. Basic emotions. *Handbook of cognition and emotion*, 98(45-60):16.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Diana G Maynard and Mark A Greenwood. 2014. Who cares about sarcastic tweets? investigating the impact of sarcasm on sentiment analysis. In *LREC 2014 Proceedings*. ELRA.

Saif M Mohammad and Peter D Turney. 2010. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 26–34. Association for Computational Linguistics.

Robert Plutchik. 1991. *The emotions*. University Press of America.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, and Prateek Vij. 2016. A deeper look into sarcastic tweets using deep convolutional neural networks. *arXiv preprint arXiv:1610.08815*.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. MELD: A multimodal multi-party dataset for emotion recognition in conversations. *CoRR*, abs/1810.02508.

Tomáš Ptáček, Ivan Habernal, and Jun Hong. 2014. Sarcasm detection on Czech and English twitter. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Association for Computational Linguistics.

Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Gerard Salton and Chris Buckley. 1987. Term weighting approaches in automatic text retrieval. Technical report, Cornell University.

Samarth Tripathi, Sarthak Tripathi, and Homayoon Beigi. 2018. Multi-modal emotion recognition on iemocap dataset using deep learning. *arXiv preprint arXiv:1804.05788*.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246.