

A Computational Model to Understand Emotions in Sarcasm

Anonymous NAACL-HLT 2021 submission

Abstract

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. In general, sarcasm is associated with a negative emotion. The question is *which negative emotion- anger, sadness, frustration, disgust, or any other?*. Previous works have extensively studied sentiment and emotion in language, while the relationship between sarcasm and emotion has been largely unaddressed. We used recently released MUsTARD dataset pre-annotated with 9 emotions, and annotated it with arousal and valence levels. Arousal and valence are important to understand the degree of emotion that led the speaker to use such figurative language. Experimental results show that our multimodal fusion models outperform the existing state-of-art systems in terms of emotion recognition. Exhaustive experimentation with each features in a modality and modality combinations is performed for both emotion and arousal-valence prediction.

1 Introduction

Emotion understanding leads to a deeper insight to the intent of the speaker. Detecting emotions and sarcasm is crucial for all services involving human interactions, such as chatbots, e-commerce, e-tourism and several other businesses. Sarcasm is a very sophisticated linguistic articulation where the sentential meaning is often disbelieved due to the linguistic incongruities. While incongruity is the key element of sarcasm, the intent could be to appear humorous, ridicule someone, or express contempt. Thus sarcasm is considered a very nuanced, or intelligent language construct which poses several challenges to emotion recognition as emotion could be completely flipped due to presence of sarcasm. Sarcasm often relies on verbal and

non-verbal cues (pitch, tone, emphasis in speech and body language in video). Even for humans, annotating the underlying emotion is challenging without the audio/video or the context of the conversation. In this paper, we aim to understand the exact emotion behind a sarcastic utterance. Since MUsTARD (Castro et al., 2019) is the only multimodal sarcastic dataset available and has only 345 sarcastic utterances, we show zero-shot emotion recognition, while training models on other existing non-sarcastic datasets. The strength of an emotion can be assessed by measuring valence and arousal, valence indicating the extent to which the emotion is positive or negative, and arousal measuring the intensity of the emotion associated (Cowie and Cornelius, 2003).

2 Related Work

Previous works have extensively studied sentiment and emotion in language, while the relationship between emotion and sarcasm has been largely unaddressed. Most of the existing research has focused on detection of sarcasm (Joshi et al., 2016, 2018). Research studying the impact of sarcasm on sentiment analysis (Maynard and Greenwood, 2014) showed that sarcasm often has a negative sentiment, but the associated emotion(s) is important to frame the response and followup communication.

In Chauhan et al. (2020), authors annotated the MUsTARD dataset with emotions and sentiment, and showed that in a multi-task setting, the primary task for sarcasm detection yielded better results with the help of secondary tasks of emotion and sentiment analysis. Since our study purely focuses on the understanding the speaker’s emotion while using sarcasm, we used their annotated basic emotions, as well as annotate the dataset with arousal and valence to understand the degree of emotion. The arousal valence annotations had 3 independent linguists as annotators with an inter-annotator agreement of 73% (Kappa score).

3 Dataset

While there exist a few data sets for sarcasm detection (Riloff et al., 2013; Ptáček et al.), (Chauhan et al., 2020) annotated the first multimodal sarcasm detection dataset MUSTARD (Castro et al., 2019) for emotions (anger, sadness, happy, neutral, frustrated, anticipation, surprise, disgust and fear). This data contains 345 sarcastic, and 345 non-sarcastic video utterances, each utterance having one or two contextual videos, which were considered by annotators while annotating. MUSTARD is a subset of Multimodal Emotion Lines Dataset (MELD)(Poria et al., 2018) which is the multimodal extension of EmotionLines dataset (Chen et al., 2018). MELD contains about 13,000 utterances from English TV-series, 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 same series and private Facebook messenger dialogues. However, both MELD and EmotionLines did not have sarcasm labels. In this paper, we used IEMOCAP (Busso et al., 2008) as a multimodal emotion labeled dataset for pretraining each of our networks. IEMOCAP has 9 emotions labeled, which are the same labels used by (Chauhan et al., 2020) annotations of MUSTARD. We didnt use CMU-MOSEI (Zadeh et al., 2018) for pretraining as the the CMU-MOSEI dataset is labelled with 6 emotions and the number of high confidence annotations is 40% of the total data. We use MELD for finetuning the networks. Since 50% of the data in MELD belongs to Neutral, we used 600 neutral samples, and all samples from rest of the classes in our finetuning phase.

4 Proposed Methodology

Since sarcasm is expressed using several non-verbal cues, we utilized the audio, video and text modalities of MUSTARD data for emotion understanding in sarcastic utterances. For all three modalities we pretrain deep self-supervised models and perform zero-shot prediction of emotion in MUSTARD. For fusion of the modalities, we used 2 layers of attention, one attention layer over each feature within a modality, and one attention layer over modalities. The aim of using multiple attention layers is to establish the relationship and importance of feature vectors obtained from the different modalities for emotion recognition and arousal-valence prediction.

4.1 Text Modality

For the text data, we obtained pretrained BERT (Devlin et al., 2018a) word embeddings for every utterance using the BERT-Base model to get a unique utterance representation of size 768. We finetuned the network on IEMOCAP (?) and MELD (Poria et al., 2018). We fine-tuned for 15 epochs using AdamW optimizer(Loshchilov and Hutter, 2017) with last 4 layers of the transformer freezed during finetuning. At test, we perform zero-shot emotion recognition on sarcastic utterances. The BERT model for emotion recognition and arousal-valence prediction is mostly same, except the last layer which for the emotion recognition problem is a 9-class classification problem, while arousal-valence is a regression problem. For comparison we trained several other models with different learned embeddings but BERT outperformed all of them.

4.2 Audio Modality

For audio, we used vocal separation to clean audio data to remove background noise and canned laughter. For our experiments, we finetuned self-supervised state-of-the-art wav2vec2.0 network (Baevski et al., 2020) which learns the latent representation by masking the spans encoded via multi-layer convolutional neural network on librispeech audio corpus (Panayotov et al., 2015) enabling it to learn generalized audio features. We finetuned this network on IEMOCAP audio and MELD audio and then tested it on full MUSTARD dataset. For comparison we also trained and tested another popular multi-task self-supervised PASE+ model (Pascual et al.)(Ravanelli et al.) with the same data, but wav2vec2.0 outperforms PASE+ marginally. Wav2vec2.0 uses contrastive loss (Oord et al., 2018) and masked language modelling objective similar to BERT (Devlin et al., 2018b) as compared to a multi-task objective in PASE+, which helps in focusing on the prosodic features in the finetuning. For baseline audio experiments, we computed low-level features such as MFCC (Mel-frequency Cepstral Coefficient), spectrogram and prosodic features and used them to show detailed ablation study on importance of each feature for emotion recognition.

4.3 Video Modality

For the video modality, we used deep residual network ResNet-18 (He et al., 2016) which tackles training issues by introducing identity skip-layer

connections that ensures that deeper network’s training error cant be larger than its shallow counterparts. We used IEMOCAP and MELD for fine-tuning and tested directly on full MUSTARD. Since the results of RESNET-18 and RESNET-154 were comparable, we continued with RESNET-18 due to its faster training.

4.4 Multimodal Fusion

We used learned input representation from networks trained on each modality through a fully connected layer and then to an inter-modality attention layer. We used an intra-modality attention only for audio to understand the relative importance of each feature as the audio model learns several deep and low-level features such as MFCC, prosodic, spectrograms etc.

5 Results

Table 1 shows the results of our zero-shot multimodal fusion model in comparison with (Chauhan et al., 2020). In (Chauhan et al., 2020), authors have used k-fold cross validation and tested using one-vs-rest strategy. We used one-vs-rest but could outperform them without including any MUSTARD samples in our training, due to use of deep semi-supervised models and similar conversational datasets for training. Although by using one-vs-rest, the accuracy is very high even for classes with very few samples, the model does not learn to predict exact emotion correctly for classes with very few samples such as fear, disgust, or surprise. Thus a multi-class classification is a better approach to measure the model’s ability to predict the exact emotion in sarcastic sentences.

Since MUSTARD is the only dataset with sarcasm and emotion and has only 345 sarcastic utterances, we show zero-shot emotion recognition on sarcastic utterances, while models are trained on non-sarcastic conversational datasets. We did perform some finetuning experiments with a subset of the sarcastic utterance, but that leads in a drop in overall precision, recall, Fscore due to the variability and insignificant examples of each class of emotion in the small sarcastic dataset. Table 2 shows the results of each modality and all combination modalities for emotion recognition task.

5.1 Error Analysis

Based on the error analysis on the ablation studies, we saw the audio model was performing better

than the text and video models alone. However, in the audio model’s confusion matrix, we observe confusion among happy and sad class, and frustration and neutral class. Since happy and sad are contrastive emotions, we performed in-depth error analysis and observed that for the misclassified happy audio segments, the spectrogram is very similar to sad audio segments. We calculated the word-overlap between happy and sad on the text modality and saw no significant adjective overlap. Therefore, it is expected that they should not be confusing classes in text modality which is reassured by the BERT model’s confusion matrix. In the multimodal experiment, we saw the confusion between happy and sad got completely eliminated, while the confusion between frustration and neutral is significantly reduced but not eliminated.

5.2 Modality-wise feature importance in Fusion

We observed that audio modality gave best results when learned in isolation and during multimodal fusion. In terms of importance of modality based on the attention scores, audio got highest importance followed by text and then video. Intra-modality attention helped us understand that in audio, MFCC features were most important for emotion classification followed by prosodic and spectrograms. Within prosodic features, loudness was the most significant feature for Sadness, Anger, Anticipation, Frustration, Happy and neutral (in order of significance). For arousal-valence predictions, prosodic features were most important followed by MFCC in audio. Text features contributed more than the spectrograms and the video features. In prosodic features, loudness followed by harmonics-to-noise ratio were the most important features before F0 and voicing features of the audio signal.

6 Conclusions

This paper provides a mechanism to predict exact emotion in sarcasm by using inter-modality attention between text, audio and video modalities in zero-shot setting. We predict basic emotions, arousal and valence to understand the intensity and polarity of the associated emotion. Although the individual components such as BERT or wav2vec or RESNET have been used before, the method of integrating known components for a *challenging problem with very limited resources* is the key take-away of this paper.

Emotion	Our Proposed			ACL2020(Chauhan et al., 2020)		
	Precision	Recall	Fscore	Precision	Recall	Fscore
Anger	82	69	74	74	85	79
Happy	79	84.8	77.6	67	79	71
Sad	66	62	64	68	82.3	74.5
Neutral	64.8	67	65.3	60.9	71.6	60.5
Frustrated	96	98	97	84.2	91.7	87.8
Anticipation	82	69	74	94	97	96.1
Surprise	93	95	94	91	95.8	93.7
Fear	96	98	97	95	97	96
Disgust	90	92	91	89	94.3	91.6

Table 1: Comparison of Zero-shot Emotion Recognition on MUSTARD dataset using one-vs-rest. Only difference is we used the full MUSTARD as test and could still outperform the state-of-art, while in (Chauhan et al., 2020) authors used k-fold cross validation thereby training their system on part of the MUSTARD data. Since in one-vs-rest, correct predictions in rest class increases the fscore, the classes with very few samples (surprise, fear, disgust) also have a high score for both systems, although the system does not perform well in predicting the exact emotion(s). **But our system outperforms significantly on classes with more samples (anger, happy, neutral and frustrated - main sarcastic emotions) by predicting the exact emotion correctly due to the feature learning of BERT and wav2vec.**

Emotion	Text			Audio			Text+Audio			Video			T + A + V		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Anger	0.16	0.06	0.09	0.33	0.28	0.30	0.42	0.29	0.34	0.16	0.63	0.26	0.19	0.06	0.09
Sad	0.35	0.14	0.20	0.37	0.44	0.40	0.37	0.71	0.49	0.00	0.00	0.00	0.21	0.13	0.16
Happy	0.34	0.09	0.10	0.38	0.17	0.24	0.00	0.00	0.00	0.00	0.00	0.00	0.13	0.06	0.08
Neutral	0.28	0.35	0.31	0.43	0.65	0.52	0.49	0.55	0.52	0.40	0.29	0.34	0.25	0.01	0.02
Fru	0.10	0.22	0.14	0.36	0.05	0.13	0.41	0.26	0.32	0.09	0.06	0.07	0.08	0.52	0.13
Ant	0.05	0.25	0.09	0.00	0.00	0.00	0.44	0.26	0.32	0.01	0.06	0.02	0.04	0.19	0.06
Surprise	0.30	0.22	0.21	0.38	0.40	0.36	0.39	0.42	0.37	0.16	0.19	0.16	0.19	0.10	0.08
Fear	0.30	0.22	0.21	0.38	0.40	0.36	0.39	0.42	0.37	0.16	0.19	0.16	0.19	0.10	0.08
Disgust	0.30	0.22	0.21	0.38	0.40	0.36	0.39	0.42	0.37	0.16	0.19	0.16	0.19	0.10	0.08
Acc	20.75%			26.99%			34.01%			19.29%			10.15%		

Table 2: Results of Zero-shot Multiclass Emotion Classification on MUSTARD for all modalities and their combinations. Results show wav2vec audio models best capture features for emotion, and audio-text model resulted in best results. (Note: P = precision, R= recall, F1= fscore, Fru= Frustrated, Ant= Anticipation, Acc= Accuracy of model)

		Text		Audio		T+A		T+A+V (W/O - S)		T+A+V	
Dimension	Test Set	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Arousal	Overall	0.186	0.36	0.19	0.36	0.14	0.28	0.26	0.37	0.30	0.43
	Non-Sarc	0.18	0.35	0.20	0.36	0.11	0.22	0.15	0.28	0.19	0.34
	Sarcastic	0.183	0.36	0.18	0.368	0.19	0.35	0.36	0.46	0.43	0.51
Valence	Overall	0.20	0.37	0.20	0.37	0.18	0.37	0.17	0.36	0.12	0.29
	Non-Sarc	0.24	0.41	0.24	0.47	0.19	0.39	0.17	0.36	0.11	0.27
	Sarcastic	0.16	0.32	0.16	0.32	0.18	0.35	0.18	0.36	0.13	0.29

Table 3: Zero-shot Arousal and Valence prediction results on MUSTARD in terms of Mean-squared error (MSE) and Mean-Average Error (MAE). Arousal prediction is slightly easier in non-sarcastic sentences as expected. Valence prediction for sarcastic utterances is observed to be easier than non-sarcastic sentences due to high variability in valence of non-sarcastic sentences (both positive and negative values, versus low variability for sarcastic valence (mostly negative)).

Ethical Considerations

This work can be deployed in real systems such as chatbots in e-commerce or other businesses wherein customer experience is of critical importance to both understand their emotions and respond accordingly. If the system works fine, it benefits the industry using it, but if it misclassifies it doesn't harm the user or the company as misclassification would not mean that the bot can reply harshly, and is equivalent of not having an emotion recognizer. The inference is real-time thus no data need to be stored and there is no potential misuse or harm from this system.

References

Alexei Baevski, Yuhao Zhou, Abdel-rahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeanette 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*.

Dushyant Singh Chauhan, SR Dhanush, Asif Ekbali, and Pushpak Bhattacharyya. 2020. Sentiment and emotion help sarcasm? a multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4351–4360.

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.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018b. Bert: Pre-training of

deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Aditya Joshi, Pushpak Bhattacharyya, and Mark J. Carman. 2018. *Investigations in Computational Sarcasm*.

Aditya Joshi, Pushpak Bhattacharyya, and Mark James Carman. 2016. Automatic sarcasm detection: A survey. *CoRR*, abs/1602.03426.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

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*.

A. Musolff. 2017. Metaphor, irony and sarcasm in public discourse. *Journal of Pragmatics*, 109:95–104.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Silviu Oprea and Walid Magdy. 2020. iSarcasm: A dataset of intended sarcasm. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1279–1289.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Santiago Pascual, Mirco Ravanelli, Joan Serrà, Antonio Bonafonte, and Yoshua Bengio. Learning Problem-Agnostic Speech Representations from Multiple Self-Supervised Tasks. In *INTERSPEECH, 2019*, pages 161–165.

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. Sarcasm detection on Czech and English twitter. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*.

Mirco Ravanelli, Jianyuan Zhong, Santiago Pascual, Pawel Swietojanski, Joao Monteiro, Jan Trmal, and Yoshua Bengio. Multi-task self-supervised learning for robust speech recognition. In *(ICASSP 2020)IEEE International Conference on Acoustics, Speech and Signal Processing*.

- 385 Mirco Ravanelli, Jianyuan Zhong, Santiago Pas-
386 cual, Pawel Swietojanski, Joao Monteiro, Jan Tr-
387 mal, and Yoshua Bengio. 2020. Multi-task self-
388 supervised learning for Robust Speech Recognition.
389 *ArXiv:2001.09239*.
- 390 Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra
391 De Silva, Nathan Gilbert, and Ruihong Huang. 2013.
392 Sarcasm as contrast between a positive sentiment
393 and negative situation. In *Proceedings of the 2013*
394 *Conference on Empirical Methods in Natural Lan-*
395 *guage Processing (EMNLP)*.
- 396 Steffen Schneider, Alexei Baevski, Ronan Collobert,
397 and Michael Auli. 2019. [wav2vec: Unsupervised](#)
398 [pre-training for speech recognition](#).
- 399 Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman,
400 Eric Nyberg, and Chris Dyer. 2014. Metaphor detec-
401 tion with cross-lingual model transfer. In *Proceed-*
402 *ings of the 52nd Annual Meeting of the Association*
403 *for Computational Linguistics*, pages 248–258.
- 404 AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Po-
405 ria, Erik Cambria, and Louis-Philippe Morency.
406 2018. Multimodal language analysis in the wild:
407 Cmu-mosei dataset and interpretable dynamic fu-
408 sion graph. In *ACL 2018*, pages 2236–2246.