A Perfect Synthesis of Multimodal Techniques for Sarcasm Detection

Abstract

A combination of verbal and non-verbal clues, such as tone changes, pronounced with greater emphasis, longer syllables, or a neutral facial expression, are frequently used to indicate sarcasm. Although textual data has been the primary focus of most previous sarcasm detection research, this work suggests that incorporating multimodal cues might greatly increase sarcasm classification accuracy. We offer the Multimodal Sarcasm Detection Dataset (MUStARD), derived from well-known TV series, to aid in the development of such systems. The dataset includes audiovisual statements that have been labeled with sarcasm together with contextual language that offers more details about the situations in which the sarcasm is used. Our first findings show that multimodal information can decrease sarcasm detection mistakes in F-score by as much as 12.9% when compared to models that only use individual modalities. The dataset is openlyaccessible to the public.

Introduction

*Sarcasm is an essential component of everyday discourse, allowing people to express scorn or contempt [1] with ease. It's usually communicated with sarcasm that has a dark undertone. As in the phrase, Perhaps it's fortunate that we arrived here. The speaker is obviously sarcastic; although framing the experience as great, their genuine meaning is negative. It's like a lesson in what not to do. But sometimes sarcasm doesn't have clear verbal indicators, so you have to look for other signs to figure out what the speaker really intends to say.

A combination of verbal and nonverbal cues, such as tone changes, emphasized words, prolonged syllables, or even a neutral facial expression, are frequently used in sarcasm. Thus, sarcasm detection necessitates seeing linguistic or contextual contradictions that can only be made sense of with more knowledge. The conversational context preceding the sarcastic remark or several modalities (Schifanella et al., 2016; Mishra et al., 2016a) may yield this further information.

This study explores the role that dialogue history and multimodal clues have in sarcasm detection. It also presents an important tool meant to propel this field's research forward. The study specifically makes the following contributions: (1) We introduce a new dataset, MUStARD (MUltimodal SARcasm Dataset), with the goal of improving multimodal sarcasm research. It includes high-quality annotations reflecting both conversational and multimodal features. (2) We illustrate the importance of a multimodal approach to addressing this problem by highlighting numerous examples where sarcasm inconsistencies span multiple modalities. (3) We introduce several benchmark models and show that multimodal systems perform significantly better than their unimodal counterparts. (4) We include the previous dialogue turns as contextual information, creating a new sub-task in sarcasm detection within conversational frameworks.

The rest of this essay is organized as follows: In Section 2, previous research on unimodal and multimodal methods for sarcasm detection is reviewed. The procedure of creating the dataset, its annotation, and the several sarcastic scenarios that are included in it are all described in Section 3. The topic of feature extraction across several modalities is covered in Section 4. Section 6 offers a thorough analysis, whereas Section 5 describes the tests carried out on the dataset. In conclusion, Section 7 offers some observations and possible obstacles for further study of this resource.

Related Work

.The automatic identification of sarcasm has attracted a lot of interest lately, mainly because of its application to sentiment analysis and human-computer interaction. Sarcasm has been studied as a sophisticated language strategy in a variety of contexts, such as voice, writing, and visual clues.

Sarcasm in Text: There have been several approaches used in the past to detect sarcasm in texts. These include rule-based approaches (Veale and Hao, 2010) and the use of lexical and pragmatic cues (Carvalho et al., 2009). Other aspects that have been studied by researchers include situational incongruity (Riloff et al., 2013), stylistic characteristics (Davidov et al., 2010), and even user-contributed tags like hashtags (Liebrecht et al., 2013).

Twitter and other social media platforms are the main sources of data in this field. There are two main approaches to annotation: remote supervision using hashtags (Davidov et al., 2010; Abercrombie and Hovy, 2016) and manual labeling (Riloff et al., 2013; Joshi et al.,2016a). A number of contextual features, such as user background (Rajadesingan et al., 2015), speaker sentiment embeddings (Poria et al., 2016), and community interactions (Wallace et al., 2015), have been studied in further detail in relation to the role of shared knowledge between the speaker and the listener (Wallace et al., 2014; Bamman and Smith, 2015). We leverage conversational structure in this dataset by combining speaker identities and previous utterances, a novel technique to dialogue-based sarcasm detection that has not previously received much attention.

Sarcasm in Speech: Recognizing prosodic cues—acoustic patterns that denote a sarcastic tone—is a major factor in sarcasm detection in speech. Early research on the subject (Cheang and Pell, 2008) found that characteristics like speech pace, harmonics-to-noise ratio, and amplitude range were indicative of sarcasm. One of the earliest researchers on sarcastic speech was Rockwell (2000), who identified slower speech rates and more intensity as possible indicators. Subsequently, Tepperman et al. (2006) investigated prosodic and spectral aspects, looking at sarcasm in and out of context. Prosodic characteristics like as stress and intonation are generally thought to be important in determining sarcasm (Bryant, 2010; Woodland and Voyer, 2011). By adding these speech attributes to our dataset, our study enables a more in-depth analysis of sarcastic vocal patterns.

Multimodal Sarcasm: Information from other modalities, in addition to text and speech, can improve sarcasm identification. These multimodal signals offer patterns that contrast or complement one another, aiding in the deciphering of sarcasm. In the past, research has frequently concentrated on how readers understand irony by combining text with cognitive cues such eye movements (Mishra et al.,2016a,b, 2017) or EEG/MEG data (Filik et al., 2014; Thompson et al., 2016). However, multimodal sarcastic expressions from the speaker's point of view have received little consideration.

Attardo et al. (2003) made one of the first investigations into this field by examining the phonological and visual indicators of sarcasm. The way in which these modalities interact was not examined in this study, though. Schifanella et al. (2016) more recently put out a multimodal approach through the analysis of visual components that go along with text in satirical internet posts. Using previously trained algorithms, they retrieved semantic visual elements from photos and combined them with textual data. Expanding upon previous initiatives, our study focuses on sarcasm in conversational discussions using videos. As far as we are aware, this is the first study to suggest a tool for video-level sarcasm identification.While Joshi et al. (2016b) presented a comparable dataset based on the television series Friends, it does not use a multimodal method and only contains textual data. We also address a number of sarcasm detection difficulties that need multimodal learning and provide a strong evaluation framework that can be expanded upon by future studies.

Dataset

We offer a novel dataset called MUStARD that consists of brief video clips that have been manually annotated for sarcasm in order to support the research on multimodal sarcasm detection.

3.1 Information Gathering The primary method of gathering data involved searching the internet, especially YouTube, for instances of possible sarcasm. We used search phrases like "sarcasm in TV shows," "Sarcasm 101," "Friends sarcasm," and "Chandler sarcasm."This produced a playlist of videos from three well-known TV shows: Sarcasmaholics Anonymous, The Golden Girls, and Friends.Our search attempts at this point were limited to satirical content.

We selected 400 films from the MELD dataset, a multimodal emotion recognition corpus that was first gathered by Poria et al. (2018) and is also based on the *Friends* series, in order to provide non-sarcastic examples. We also added snippets from The Big Bang Theory, a program frequently linked to sardonic humor, to our dataset. Using the laughter of the live studio audience as a gauge, episodes from the first eight seasons were divided apart. We identified segment boundaries using open-source laughter detection software (Ryokai et al., 2018). Subtitle timestamps were then used to refine the boundaries.
After our gathering operation, we had a dataset with 6,421 videos. While some videos had already been labeled for sarcasm, the majority did not, requiring a laborious manual annotation procedure.

3.2 Process of Annotation We created a unique web-based annotation tool that allowed commentators to assess the degree of sarcasm in the material by showing each video next to its transcript. Annotators had the option through the interface to mark videos that had technical problems, such audio and video out of sync. Furthermore, when annotators felt it was important to do so in order to form appropriate assessments, they might observe the previous conversational background. Because there were so many videos, our web tool was used to annotate them in batches of four videos at a time (Figure 2). There were two separate stages to the annotating process.

Since The Big Bang Theory videos made up the majority of the collection, we focused on them throughout the first phase.Independently, two graduate students who had previously studied sarcasm annotated 5,884 utterances. At first, a startling 98% of them were classified as non-sarcastic, which led to a low level of agreement between annotators (Kappa score: 0.1463). Seeing this problem, we stopped the process and had a discussion to work out the differences. After reviewing twenty movies that were chosen at random, the two annotators resolved any discrepancies in their labels and annotated the dataset again.

The result was a much higher Kappa score of 0.2326. A third annotator examined the disputed films and decided on the proper label after resolving the remaining disagreements.

We turned our attention to 624 videos from Friends, The Golden Girls, and Sarcasmaholics Anonymous in the second round. The two annotators tagged each video separately, just like in the first phase. With a Kappa score of 0.5877, the inter-annotator agreement was significantly greater this time. Once more, a third annotator settled disagreements. 6,365 movies make up the final annotated dataset, 345 of which are classified as sarcastic and 6,020 as non-sarcastic. This well-balanced dataset provides a solid foundation for next multimodal sarcasm detection studies.

Transcriptions

Overview of the Dataset and Transcriptions Some of the video clips, especially those from The Big Bang Theory and the MELD dataset, already had transcripts or subtitles accessible due to the variety of sources from which the clips were gathered for our study. While the transcripts from The Big Bang Theory were recovered through manual substring matching with episode subtitles, we made direct use of the MELD transcriptions. To assure accuracy, we manually transcribed the remaining video clips.

3.4 MUStARD Sarcasm Dataset

We created a balanced sample set inside our dataset, MUStARD, in order to plan experiments targeted at capturing the multimodal features of sarcasm detection. From the 6,365 initially annotated clips, a thorough selection process was used to produce this balanced dataset, which includes an equal amount of sardonic and non-sarcastic videos. We started by choosing every video that has been classified as sarcastic. We next randomly selected an equivalent number of clips from the non-sarcastic subset, giving preference to those that had been examined by several annotators in order to increase reliability.

There are 690 videos in the final collection, with an equal number of sarcastic and non-sarcastic samples. Detailed statistics on the source material, character representation, and label distribution are shown in Figures 3 and 4. Every video in our dataset is referred to as a utterance for the sake of this investigation. In order to provide a more comprehensive portrayal of a character's speech, we define an utterance to include several consecutive sentences said by the same person. Consequently, single sentences make up 61.3% of the utterances, with numerous sentences making up the remaining portion.

Every utterance is accompanied by context utterances, representing the dialogue turns preceding the target utterance, spoken by other characters involved in the conversation. In some cases, the context involves multi-party dialogues. The length of the context is manually adjusted to ensure that it provides coherent background information for the target utterance. Table 1 offers a comprehensive overview of the dataset's utterances.

Statistics	Utterance	Context
Unique words	1991	3205
Avg. utterance length (tokens)	14	10
Max. utterance length (tokens)	73	71
Avg. duration (seconds)	5.22	13.95

Table 1: Dataset statistics by utterance and context

Three primary modalities are used to represent each utterance and its context: text transcription, audio, and video. Speaker identities are also included with each speech, which aids with character differentiation throughout the dataset. A sarcastic statement and its context are shown in Figure 1. The distribution of labels and the primary characters in the dataset are shown in detail in Figures 4a and 4b. Chandler Bing and Sheldon Cooper, two well-known figures in caustic humor, are particularly well-represented. We made sure that non-sarcastic samples for these characters were also included in order to reduce any potential biases resulting from them. Furthermore, minor roles were included to further investigate speaker-specific sarcasm recognition patterns, such as Dorothy from The Golden Girls, who maintains a continuous sarcastic tone.

3.5 Elaborate Views Multimodal information is crucial because it provides clues beyond the written word that are frequently needed to identify sarcasm in textual content. Here, we look at particular examples from the MUStARD dataset to show how important it is to use various modalities when sarcasm detection is needed.

Multimodality's Role: Two instances of sarcasm resulting from differences between modalities are shown in Figure 5. In the first case, the speaker's facial emotions do not match the language's suggestion of fear or wrath. In the second example, the speaker's vocal tone and facial expressions communicate disinterest, despite the words seeming to be a complement. The mismatch between voice, text, and facial expressions in both situations turns into a crucial clue to sarcasm.

The identification of sarcasm also heavily relies on multimodal information, such as vocal tone. For instance, even though a sentence appears neutral in writing, the speaker's tone may indicate sarcasm when it is heard. The speaker's intention can be conveyed through sarcastic vocal tones that range from self-deprecating to overly theatrical or irritating. Another way to recognize sarcasm is to put too much emphasis on a particular word. Think about the statement, "You did really well." It's obvious that this is sarcastic when too much emphasis is put on "really." Examples from the dataset where vocal accent indicates sarcasm are shown in Figure 6. It's crucial to remember that sarcasm doesn't necessarily require signals that contradict with one

another across many modalities. Rather, combining complimentary data from several modalities improves a model's capacity to identify the nuanced and intricate signs of sarcasm.

The Significance of Context: Figure 7 illustrates two situations in which sarcasm can be reliably identified based on the conversational context. In the first example, it takes knowledge of the debate topic—tanning—to reveal a satirical allusion to the sun. Likewise, one may only comprehend the satirical reference to a Venus flytrap as a romantic partner if they have listened to the previous exchange. These illustrations show how important context is in identifying sarcastic speech.Because the MUStARD dataset contains conversational context, models can access more background data, which improves the accuracy of sarcasm detection.Future methods could make use of common sense to identify sarcastic expressions that are often ludicrous, like "dating a Venus flytrap."In conclusion, the MUStARD dataset provides a rich resource for researching sarcasm detection and opens chances for developing models that comprehend and process the subtleties of sarcasm in a more sophisticated manner by combining multimodal data with conversational context.

Multimodal Feature Extraction

Multimodal Method for Extracting Features Three major modalities that are present in the dataset are used by our system to extract various learning features. An extensive explanation of each modality's extraction procedure is provided below:

Textual Elements: To represent every utterance in the dataset and extract the substance of the textual data, we use BERT (Devlin et al., 2018). In particular, we create a dense vector representation of every speech with BERT-Base by averaging the final four transformer layers of the initial token ([CLS]); this yields a dimensionality of 768 (dt = 768). Although the pre-trained 300-dimensional GloVe word vectors were used in our experiments (Pennington et al., 2014), averaging across tokens produced worse results than BERT.

Speech Features: Important intonational and tonal signals are provided by the audio modality. Using Librosa (McFee et al., 2018), we extract low-level acoustic features to capture qualities like pitch and intonation (Tepperman et al., 2006). A heuristic vocal-extraction technique is used to decrease background noise once the audio signals are loaded at a sample rate of 22050 Hz. After processing the audio, it is divided into dw non-overlapping windows. These windows are then used to extract the spectral centroid, melspectrogram,MFCC, and their derivatives (delta). By averaging the features across windows, the audio representation is combined into a 283-dimensional (da = 283) feature set.

Visual Features: Using the pool5 layer of a ResNet-152 (He et al., 2016) model that has been pre-trained on ImageNet (Deng et al.2009), visual representations are retrieved from the video frames. Each frame is preprocessed by center-cropping, resizing, and normalizing it. The dv = 2048 dimensional vectors for each frame are averaged to provide uv = 1/f($\sum$i uvi), which is a unified representation for the full utterance and the visual feature vector. We chose to use this straightforward averaging method in order to preserve consistency across modalities,

even though more sophisticated encoding approaches such as recurrent neural networks might be utilized.

Experiment

We evaluated individual modalities as well as different combinations of the modalities present in our dataset in order to investigate the effects of multimodal integration in sarcasm detection. In addition, we investigated the possible advantages of adding speaker-specific and contextual data to improve the model's prediction skills.Setup for an Experiment Our studies included two main evaluation sets. To preserve label balance, we created each fold in the initial batch of data using a stratified randomization technique. This was done using five-fold cross-validation. One fold was utilized as the test set and the other four folds were used for training in each of the K iterations. With part of the training set, validation was carried out.A speaker-dependent experimental setup was created since there was speaker overlap between the training and testing sets as a result of the folds being randomized.We used a speaker-independent setup in the second evaluation set. Here, we made sure that the training and test sets did not contain any speech from the same speaker. To be more precise, the training set consisted of quotes from TV series including The Big Bang Theory, The Golden Girls, and Sarcasmaholics Anonymous, whereas the test set included quotes from Friends. Section 6 elaborates on the reasoning behind this configuration.Our assessment measures for these studies were F-score, precision, and recall, which were weighted based on the distribution of sardonic and non-sarcastic utterances in the class.We presented the average outcomes of the five-fold cross-validation approach for the speaker-dependent scenario.baselines We used three important baseline techniques to compare our experiments to:Majority: A straightforward baseline that categorizes all speech as falling into the non-sarcastic majority class.Random: Predictions are uniformly picked from the test set and assigned at random using this baseline.Support Vector Machines (SVM): Because SVM works well with smaller datasets, we used it as our main machine learning baseline. We utilized a scaled gamma and an RBF kernel, using the scikit- learn (Pedregosa et al., 2011) SVM implementation.For every experiment, the penalty term (C) was tweaked between values of 1, 10, 30, 500, and 1000, using it as a hyperparameter. Feature scaling was used in the speaker-dependent setup by deducting the mean and dividing the result by the standard deviation. Early fusion was applied to multimodal configurations, concatenating features from many modalities.Multimodal Classification of Sarcasm The outcomes of the speaker-dependent setup's sarcasm categorization are displayed in Table 2. With a weighted F-score of 33.3% (66.7% for non-sarcastic utterances and 0% for sarcastic utterances), the Majority baseline performed the worst. The best performance was obtained with visual features among the unimodal baselines. The model performed much better and reached the highest accuracy when we combined textual features using concatenation.The model performed much better and reached the highest accuracy when we combined textual features using concatenation. However, because of less-than-ideal audio feature extraction, adding the audio modality to the tri-modal variation marginally decreased performance. Combining textual and visual modalities produced significant benefits overall, with a 12.9% relative error reduction over the unimodal versions.The utterances where the unimodal textual model failed to identify sarcasm were manually evaluated. These were identified by the bimodal model (text and visual). For accurate recognition in the majority of these cases, additional multimodal signals were needed because sarcastic cues were not always evident in the text alone (refer to Fig. 9).Since the model was unable to learn speaker-specific patterns, the speaker- independent arrangement proved more difficult than the speaker-dependent setup.More generalization was required when unknown speakers were included to the test set. Furthermore, this configuration produced a novel setting for every modality, which made it a perfect testing ground for multimodal sarcasm detection studies. The increased complexity was also evident in the model's training procedure, which called for finer tuning (higher C values) in order to get the SVM to perform satisfactorily on tests.Baseline performances under the

speaker-independent arrangement are shown in Table 3. In this instance, the multimodal arrangements did not perform appreciably better than the unimodal ones.

Sarcasm prediction was more heavily influenced by the auditory modality than by the speaker-dependent outcomes. Performance somewhat increased when textual characteristics were paired with audio. Examining the correctly predicted sarcastic utterances (recognized by the text-plus-audio model but missed by the text-only model), we found that, in agreement with previous studies by Attardo et al. (2003), the correctly categorized samples tended to have a higher mean pitch. On the other hand, inaccurate guesses frequently showed considerable pitch variability, suggesting that temporal patterns in the audio channel should be the main focus of future study.Remarkably, video functions did not operate as well in this configuration.Remarkably, video functions did not operate as well in this configuration. We speculate that there might be biases toward particular characters due to the model's limited complexity and that the object-based visual elements that were recovered might not be sarcasm-specific. Statistics from Fig. 10, which are covered in the following section, support this. Sarcasm detection models ought to take into consideration discrepancies between a speaker's emotional content and their facial expressions, according to additional examination of the best model's mistakes.The Significance of Speaker Information and Context We also explored whether sarcasm predictions might be improved by adding extra context (e.g., previous utterances) and speaker identification.Speaker attributes were represented by one-hot encoding vectors, while contextual features were produced by averaging the embeddings of previous utterances.Table 4 presents the comparison between the best multimodal model and the textual baseline for both evaluation settings. When context features were introduced to the text-plus-audio model in the speaker-independent configuration, we saw a minor improvement; however, this trend did not hold true for the other models. This might be explained by the fact that averaging representations across the course of a conversation results in the loss of temporal information.Because the model made use of speaker-specific patterns, the inclusion of speaker features enhanced performance in the speaker-dependent setup for the textual modality. However, there was little gain in the best multimodal option (text + video), indicating that the input characteristics already implicitly collected speaker-specific information.Because there was no speaker overlap between the training and test sets, adding speaker information to the speaker-independent setup did not result in performance gains as anticipated.

Algorithm	Modality	Precision	Recall	F-Score
Random	-	49.5	49.5	49.8
SVM	TA V	65.1 65.9 68.1	64.6 64.6 67.4	64.6 64.6 67.4
	T+A T+V A+V T+A+V	66.6 **72.0** 66.2 71.9	66.2 **71.6** 65.7 71.4	66.2 **71.6** 65.7 71.5
$\Delta_{multi-unimodal}$ Error rate reduction		↑↑ 3.9% 12.2%	↑↑ 4.2% 12.9%	↑↑ 4.2% 12.9%

Table 2: Speaker-dependent setup. All results are aver- aged across five folds where each fold present weighted F-score across both sarcastic and non-sarcastic classes.

Random	-	51.1	50.2	50.4
	T A V	60.9 **65.1** 54.9	59.6 62.6 53.4	59.8 62.7 53.6
	T+A	64.7	**62.9**	**63.1**

SVM	T+V A+V T+A+V	62.2 64.1 64.3	61.5 61.8 62.6	61.7 61.9 62.8
$\Delta_{multi-unimodal}$		0.4%	0.3%	0.4%
Error rate reduction		↓ 1.1%	↑ 0.8%	↑ 1.1%

Conclusion and Future Work

We conducted an extensive investigation into the field of multimodal learning for sarcasm detection in this work, and we introduced a new dataset, MUStARD, which contains both sarcastic and non-sarcastic examples from various video sources. Through specific examples
drawn from this dataset, we were able to demonstrate how important multimodal analysis is for correctly identifying sarcasm. Then, we suggested models that may combine the effectiveness of three main modalities: verbal, visual, and textual cues. To further improve the accuracy of the model, we also looked into adding context and speaker-specific data.

Our baseline studies demonstrated that using many modalities is important for sarcasm detection, with multimodal techniques outperforming unimodal ones by a significant margin and reducing relative error rates by as much as 12.9%. We also discovered some important issues with our research that need to be addressed in other studies. Among them are:

1. Sophisticated Methods of Multimodal Fusion: Even while our current model depends on early fusion, more advanced techniques like Tensor Fusion or Canonical Correlation Analysis (CCA) may be investigated in the future to better capture the interplay between modalities. Developing fusion tactics that highlight the innate modalities' inconsistencies—which are frequently signs of sarcasm—could be another approach.

2. Multiparty Dialogue Analysis: Future work could benefit from models that comprehend not only the individual sarcasm cues but also the interpersonal dynamics, gestures, and facial expressions of multiple individuals in a conversation, as many conversations in our dataset involve multiple speakers

3. Improved Neural Models: While SVM models fared well on our dataset, we noticed that the small size of the dataset caused neural models, such as CNNs, to suffer from overfitting. Subsequent endeavors may concentrate on surmounting this constraint by employing strategies like pre-training, transfer learning, or low-parameter models that are more appropriate for smaller datasets.

4. Contextual Sarcasm Detection: Although the last statement in a discussion is the focus of our dataset, the earlier statements in a conversation frequently provide crucial background. The goals and intentions of the interlocutors could be included to greatly enhance the identification of sarcasm in conversational contexts.

5. Speaker Localization: By precisely dentifying the primary speaker in multiparty videos and concentrating on their movements and facial expressions, the model's capacity to recognize sarcasm may be significantly enhanced. In summary, we think that the MUStARD dataset offers a strong basis for

upcoming advancements in multimodal sarcasm detection, and resolving the issues raised here will spur more.

References

Gavin Abercrombie and Dirk Hovy. 2016. Putting sar- casm detection into context: The effects of class im- balance and manual labelling on supervised machine classification of twitter conversations. In *Proceed- ings of the ACL 2016 Student Research Workshop*, pages 107–113.

Salvatore Attardo, Jodi Eisterhold, Jennifer Hay, and Isabella Poggi. 2003. Multimodal markers of irony and sarcasm. *Humor*, 16(2):243–260.

David Bamman and Noah A Smith. 2015. Contextual- ized sarcasm detection on twitter. *ICWSM*, 2:15.

Gregory A Bryant. 2010. Prosodic contrasts in ironic speech. *Discourse Processes*, 47(7):545–566.

Evgeny Byvatov, Uli Fechner, Jens Sadowski, and Gis- bert Schneider. 2003. Comparison of support vec- tor machine and artificial neural network systems for drug/nondrug classification. *Journal of chemical in- formation and computer sciences*, 43(6):1882–1889.

Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2018. OpenPose: realtime multi- person 2D pose estimation using Part Affinity Fields.In *arXiv preprint arXiv:1812.08008*.

Paula Carvalho, Luís Sarmento, Mário J Silva, and Eu- génio De Oliveira. 2009. Clues for detecting irony in user-generated contents: oh...!! it's so easy;-. In *Pro- ceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, pages53–56. ACM.

Henry S Cheang and Marc D Pell. 2008. The sound ofsarcasm. *Speech communication*, 50(5):366–381.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Semi-supervised recognition of sarcastic sentences in twitter and amazon. In *Proceedings of the four- teenth conference on computational natural lan- guage learning*, pages 107–116. Association for Computational Linguistics.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hier- archical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages248–255. IEEE.

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

Ruth Filik, Hartmut Leuthold, Katie Wallington, and Jemma Page. 2014. Testing theories of irony pro- cessing using eye-tracking and erps. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(3):811.

Devamanyu Hazarika, Soujanya Poria, Sruthi Gorantla, Erik Cambria, Roger Zimmermann, and Rada Mihal- cea. 2018. Cascade: Contextual sarcasm detection in online discussion forums. *Proceedings of the 27th International Conference on Computational Linguis- tics*, page 1837–1848.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recog- nition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Harold Hotelling. 1936. Relations between two sets ofvariates. *Biometrika*, 28(3/4):321–377.

Aditya Joshi, Pushpak Bhattacharyya, Mark Carman, Jaya Saraswati, and Rajita Shukla. 2016a. How do cultural differences impact the quality of sarcasm annotation?: A case study of indian annotators and american text. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Her- itage, Social Sciences, and Humanities*, pages 95– 99.

Aditya Joshi, Vinita Sharma, and Pushpak Bhat- tacharyya. 2015. Harnessing context incongruity for sarcasm detection. In *Proceedings of the 53rd An- nual Meeting of the Association for Computational Linguistics and the 7th International Joint Confer- ence on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 757–762.

Aditya Joshi, Vaibhav Tripathi, Pushpak Bhat- tacharyya, and Mark J Carman. 2016b. Harnessing sequence labeling for sarcasm detection in dialogue from tv seriesfriends'. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Lan-guage Learning*, pages 146–155.

Y Alex Kolchinski and Christopher Potts. 2018. Rep- resenting social media users for sarcasm detection. In *Proceedings of the 2018 Conference on Empiri- cal Methods in Natural Language Processing*, pages 1115–1121.

CC Liebrecht, FA Kunneman, and APJ van Den Bosch. 2013. The perfect solution for detecting sarcasm in tweets# not. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Senti- ment and Social Media Analysis*, pages 29–37. NewBrunswick, NJ: ACL.

Navonil Majumder, Soujanya Poria, Devamanyu Haz- arika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2018. Dialoguernn: An attentive rnn for emotion detection in conversations. *arXiv preprint arXiv:1811.00405*.

Brian McFee, Matt McVicar, Stefan Balke, Carl Thomé, Vincent Lostanlen, Colin Raffel, Dana Lee, Oriol Nieto, Eric Battenberg, Dan Ellis, Ryuichi Yamamoto, Josh Moore, WZY, Rachel Bit- tner, Keunwoo Choi, Pius Friesch, Fabian-Robert

Stöter, Matt Vollrath, Siddhartha Kumar, nehz, Si- mon Waloschek, Seth, Rimvydas Naktinis, Dou- glas Repetto, Curtis "Fjord" Hawthorne, CJ Carr, João Felipe Santos, JackieWu, Erik, and Adrian Holovaty. 2018.

Abhijit Mishra, Kuntal Dey, and Pushpak Bhat- tacharyya. 2017. Learning cognitive features from gaze data for sentiment and sarcasm classification using convolutional neural network. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 377–387.

Abhijit Mishra, Diptesh Kanojia, and Pushpak Bhat- tacharyya. 2016a. Predicting readers' sarcasm understandability by modeling gaze behavior. In *AAAI*,pages 3747–3753.

Abhijit Mishra, Diptesh Kanojia, Seema Nagar, Kun- tal Dey, and Pushpak Bhattacharyya. 2016b. Har- nessing cognitive features for sarcasm detection. In *Proceedings of the 54th Annual Meeting of the As- sociation for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1095–1104.

D. Cournapeau, M. Brucher, M. Perrot, and E. Duch- esnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language process- ing (EMNLP)*, pages 1532–1543.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, and Prateek Vij. 2016. A deeper look into sarcas- tic tweets using deep convolutional neural networks. In *Proceedings of COLING 2016, the 26th Inter- national Conference on Computational Linguistics: Technical Papers*, pages 1601–1612.

Soujanya Poria, Devamanyu Hazarika, Navonil Ma- jumder, Gautam Naik, Erik Cambria, and Rada Mi- halcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019. Emotion recognition in conversation: Research challenges, datasets, and re-cent advances. *arXiv preprint arXiv:1905.02947*.

Ashwin Rajadesingan, Reza Zafarani, and Huan Liu. 2015. Sarcasm detection on twitter: A behavioral modeling approach. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 97–106. ACM.

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 Lan- guage Processing*, pages 704–714.

Patricia Rockwell. 2000. Lower, slower, louder: Vo- cal cues of sarcasm. *Journal of Psycholinguistic Re- search*, 29(5):483–495.

Kimiko Ryokai, Elena Durán López, Noura Howell, Jon Gillick, and David Bamman. 2018. Capturing, representing, and interacting with laughter. In *Pro- ceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 358. ACM.

R Schifanella, P de Juan, J Tetreault, L Cao, et al. 2016. Detecting sarcasm in multimodal social platforms. In *ACM Multimedia*, pages 1136–1145. ACM.

Joseph Tepperman, David Traum, and Shrikanth Narayanan. 2006. " yeah right": Sarcasm recogni- tion for spoken dialogue systems. In *Ninth Interna- tional Conference on Spoken Language Processing*.

Dominic Thompson, Ian G Mackenzie, Hartmut Leuthold, and Ruth Filik. 2016. Emotional re- sponses to irony and emoticons in written language:evidence from eda and facial emg. *Psychophysiol- ogy*, 53(7):1054–1062.

Tony Veale and Yanfen Hao. 2010. Detecting ironic in- tent in creative comparisons. In *ECAI*, volume 215, pages 765–770.

Byron C Wallace, Eugene Charniak, et al. 2015. Sparse, contextually informed models for irony de- tection: Exploiting user communities, entities and sentiment. In *Proceedings of the 53rd Annual Meet- ing of the Association for Computational Linguistics and the 7th International Joint Conference on Natu- ral Language Processing (Volume 1: Long Papers)*,volume 1, pages 1035–1044.

Byron C Wallace, Laura Kertz, Eugene Charniak, et al. 2014. Humans require context to infer ironic intent (so computers probably do, too). In *Proceedings of the 52nd Annual Meeting of the Association for Com- putational Linguistics (Volume 2: Short Papers)*, vol-ume 2, pages 512–516.

Silvio Amir Byron C Wallace, Hao Lyu, and Paula Car- valho Mário J Silva. 2016. Modelling context with user embeddings for sarcasm detection in social me-dia. *CoNLL 2016*, page 167.

Jennifer Woodland and Daniel Voyer. 2011. Con- text and intonation in the perception of sarcasm. *Metaphor and Symbol*, 26(3):227–239.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cam- bria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empiri- cal Methods in Natural Language Processing*, pages 1103–1114.