Multimodal Emotion Classification in Sarcastic Utterances

Apoorva Nunna, Pushpak Bhattacharyya, Anupama Ray
Department of Computer Science and Engineering, IIT Bombay,
IBM Research India
{apoorvanunna, pb}@cse.iitb.ac.in, anupamar@in.ibm.com

Abstract

Automatic emotion analysis, where the model tries to identify the underlying emotion in the input, is an area that is being actively researched upon. Sarcasm is often referred to as a tool that helps express contempt or mock someone. The intrinsic nature of 'saying something but meaning the opposite,' i.e., the incongruity, is what makes detecting the presence of sarcasm challenging. Bringing together these two worlds, that is, while it is known that the polarity associated with sarcasm is negative, pinpointing the negative emotion that the author/speaker expresses is a very challenging task. Also, the idea of performing automatic emotion analysis specifically on sarcastic utterances has been far less explored than emotion recognition in general.

Through our work, we perform the task of detecting the *implicit emotion in a given sarcastic utterance*. We utilise the previously gathered MUStARD++ dataset which contains 601 sarcastic and 601 non-sarcastic utterances where each instance is available in multiple modalities - text, audio and video. Along with detecting the discrete emotion in sarcastic utterances we also extend the dataset with the two important continuous dimensions of emotion - Valence and Arousal. We perform exhaustive experimentation with multimodal (text, audio and video) fusion models to establish a benchmark for implicit emotion recognition in sarcasm. In order to overcome the challenge of the small size of our dataset, we also attempt and observe the effect of automatic data augmentation techniques.

Detecting sarcasm not only depends on the actual content spoken, pitch, tone, facial expressions, and context of an utterance but also on personal traits like language proficiency and cognition capabilities of the reader/listener. Studies show the presence of distinctive eye movement and fixation patterns while readers try to read and understand sarcastic sentences. With this as the motivation, as part of this project, we gather eye-tracking data on our sarcastic utterances presented in a conversational format. This data can be utilized in the future to study the role of gaze in detecting conversational sarcasm. The extracted eye movement and fixation patterns can also be used to explain sarcasm, given the contextual utterances spoken in a conversation.

The literature review that was created as a part of the overall thesis effort is presented in this paper.

1 Problem Definition

Sarcasm is a linguistic phenomenon where the intended emotion is disguised by a different or even opposite emotion. The Figure 1 shows one such example with the text indicating amusement or happiness while the facial expression reveals the disgust of the speaker. We refer to the surface emotion that appears to have been conveyed by the author/speaker as 'explicit emotion' and the intended emotion is referred to as the 'implicit emotion'. Through our work, we intend to detect the implicit emotion that is conveyed by a sarcastic utterance.
Sarcasm does not always have to present itself in words and is quite often than not expressed using audio-visual cues. Due to this reason, the sarcastic utterances that we analyse and study are multimodal in nature, each associated with audio and video content in addition to text.
So given **one or more modalities of input** the model we build needs to identify which **implicit emotion class** the input belongs to. Considering the fact that sarcasm always has a negative implicit emotion (Joshi et al., 2016a) we consider 5 implicit emotion classes which are as follows - Anger, Sadness, Frustration, Ridicule and Disgust.
As part of our methodology, we also try to leverage other information regarding the instances, like speaker information, context in which the utterance was spoken; all of which are a part of the MUStARD++ dataset.

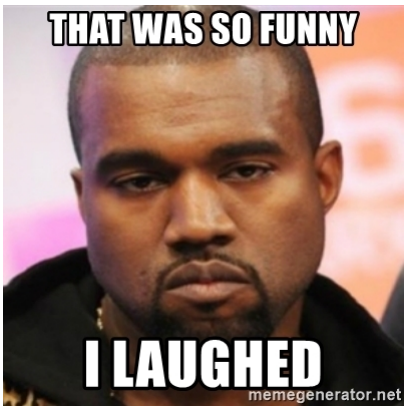

Figure 1: Incongruent emotions in sarcasm[1]

2 Motivation

2.1 The Challenge of Sarcasm

Sarcasm, according to the dictionary of Cambridge, is defined as 'the use of remarks that clearly mean the opposite of what they say, made in order to hurt someone's feelings or to criticize something in a humorous way'[2].

Similarly, Emotion Analysis essentially involves understanding the emotion contained in a piece of input. With the rise in usage of social platforms to communicate and express opinions, there are a multitude of scenarios and applications in which detecting emotion would be of significance. There are a lot of challenging problems and scope for improvement within the realm of emotion analysis in NLP making it an interesting field to study.
While both of these problems are challenging by themselves, a combination of them poses a much intriguing challenge. In other words, while detecting sarcasm itself is one challenge, once detected figuring out the intended emotion of the speaker/author is non-trivial.

2.2 Motivation for Multimodality

While incongruity lies at the heart of any sarcastic utterance, it is not necessary that this incongruity is always presented in the words. More often than not, it becomes difficult or impossible to detect the presence of sarcasm without additional cues that come with the context of the utterance or some visual or acoustic signal like rolling of eyes, increased pitch, indifference in tone etc.
Example: *Yeah, the movie was great *makes an unamused face**

[1] https://memegenerator.net/img/instances/35117460.jpg

2.3 Advantages of Emotion Analysis

The problem is also worth pursuing for improving various applications of emotion detection few of which are listed below:

- **Intelligent chat bots:** Chat bots these days are commonly being used in a variety of applications to automate communication with the end user to some extent. In scenarios where a human being interacting with the chatbot is being sarcastic, a chatbot with only good sarcasm detection algorithm, which only detects sarcasm or one with just good emotion detection algorithm which identifies the explicit emotion in the input would not suffice. An example scenario is presented in table 1

Author	Message
Chatbot	Would you like to cancel the order?
User	No, of course not, I have been ranting for the past 5 minutes about what a terrible product it is, just so I could order 1000 more of these
Chatbot	Sorry, the seller has a limit of 5 on this item. Would you like to increase the quantity to 5?

Table 1: Non-ideal Chatbot Conversation

- **Online Reviews:** Online reviews are automatically analysed and their sentiment is studied in several organizations. These days, along with sentiment a more granular analysis of emotion is becoming popular. Understanding the explicit emotion, which is quite often the opposite of the implied emotion in a sarcastic review, could result in giving entirely incorrect insights to a firm's research.
  About a wireless epilator
  Example 1: I love the fact that the battery in this drains in 15 minutes
  Example 2: Beautiful pink device that will take you to the gates of hell upon usage
  These would not count as defects in the company's analysis with a conventional sentiment/emotion analysis framework.

[2] https://dictionary.cambridge.org/dictionary/english/sarcasm

3 Literature Survey

The following chapter discusses the literature that is relevant and forms basis to the problem statement. We discuss the research work that was studied in the path of progress.

3.1 Approaches to Emotion Recognition

There are different approaches with which the problem of emotion analysis can be dealt. Broad overview of these approaches is provided in this subsection

3.2 Rule-based Approach

The most straight-forward method is the rule-based approach to solving emotion analysis problems. This indicates detecting emotions in a given input by applying certain rules and relying on the presence of certain structure or words in the input. There are broadly two important traditional techniques based on (Acheampong et al., 2020) and (Joshi et al., 2016b)

1. **Keyword Spotting**: This relies on the presence of certain affective words that indicate emotion in the input. This approach relies on Emotion Lexicons or dictionaries like WordNet Affect, NRC Emotion Lexicon, DepecheMood etc. These provide the emotion keywords to look for in a sentence, in order to associate it with a particular emotion

2. **Lexical Affinity**: This method extends over the Keyword Spotting method in the sense that, along with looking for affective items, in this method, random words are given some probabilistic affinity. However according to (Joshi et al., 2016b), this method poses two disadvantages

    - This approach may not be able to deal with negations and different word senses.
    - The probabilities assigned are usually dictated by the source of the linguistic corpora, there by making it domain dependent

In general, Rule-based traditional approaches usually require a lot more effort compared to further developed approaches. They usually end up being domain-specific and non-robust to changes.

3.2.1 Statistical Learning Approach

Statistical learning algorithms are one of the most common approaches used in emotion analysis from text and act as baselines. The machine learning algorithm is fed with a large training corpus which gives the model the ability to learn automatically from experience. They learn not only from the lexical features and valence, but also from paradigmatic features like (!,?).The most popular machine-learning algorithm used in this paradigm is the Support Vector Machine. The Figure 2 shows the main steps in SVM

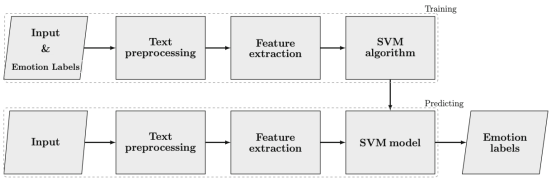

Figure 2: Overview of Steps in SVM (Alswaidan and Menai, 2020)

(Alswaidan and Menai, 2020) explains these steps with the first step being pre-processing of text from the dataset. This may include tokenization, stop words removal, lemmatization/stemming, and POS tagging. The next step would be to extract useful features followed by selecting the features with most information gain.It is to be noted that, when it comes to classical machine learning approaches, feature extraction is shown to hold a lot of importance. Given the feature set and emotion labels, the SVM algorithm outputs an optimal hyperplane. Finally, the resultant trained SVM model can be used to classify emotions in unseen text.

However, it is to be noted that the rule-based and classical machine learning approaches, often fail to generalize and capture the nuances of language effectively.

3.2.2 Deep Learning Approach

Deep Learning is a branch of Artificial Intelligence, that relies on the concept of letting computers learn from experience and understand the world in terms of a hierarchy of concepts, such that complicated concepts are learnt by building them out of simpler ones (Goodfellow et al., 2016). This approach is of great importance considering the popularity and performance being delivered by the Deep Neural Networks in various NLP problems including Emotion Analysis.

Types of Deep Neural Networks Deep Neural net-

works have undergone a rich phase of evolution, beginning with the concept of perceptron and building upon it with Multi Layer Perceptron(MLP) to finally evolving to the latest Transformers.

FeedForward Neural Networks (aka MLP) are found in several ML applications and the idea of FFNN is to define a mapping $y = f(x; \theta)$ and learn the parameters θ in order to approximate f in the best manner possible. It mainly relies on the concept of backpropagation, in order to adjust the parameter values with the aim of minimizing the loss function

The idea of capturing context while performing sequential tasks using encoder-decoder architectures found its limitations in case of capturing long-range dependencies in long sentences. The concept of **Attention** introduced by (Bahdanau et al., 2014) and (Luong et al., 2015) made a revolutionary difference in handling this problem, specially in the machine translation systems. The main idea behind attention is to enhance importance to essential parts of the input in comparison to others.

Transformers is a concept proposed by (Vaswani et al., 2017) which uses the idea of attention to boost the speed with which models are trained by allowing parallelism. These are the kind of models that are most commonly being used these days and have been used in our project. Not only do the pre-trained versions of transformers help in classification tasks, but are also quite extensively used for the step of feature extraction. Several transformer models like (Devlin et al., 2018), (Lewis et al., 2019), (Liu et al., 2019b), (Raffel et al., 2019) were used in the course of this project for feature extraction.

3.3 Sarcasm Detection

(Joshi et al., 2016a) is one of the most important papers that talks about the basics, theories, types of sarcasm. It points out the milestones done in automatic sarcasm detection, the challenges and applications of the task.

3.4 Multimodal Datasets

One of the major challenges faced in this project is quantity of data specially in a multi-modal setting. Hence an attempt was made to find more multimodal datasets to see if they could potentially be annotated to suit our problem. While none of the ready-made datasets seemed feasible enough to be annotated in the given time constraints and due to being asynchronuous with our existing dataset, we mention the datasets found in our survey here, since it might be useful for some other inspired project.

3.4.1 IEMOCAP

Interactive Emotional Dyadic Motion Capture (IEMOCAP)[3] database is an acted, multimodal and multispeaker database, collected at SAIL lab at USC. It contains approximately 12 hours of audiovisual data, including video, speech, motion capture of face, text transcriptions. It contains sessions with mostly 2 actors performing scripts that focused on expressing emotions. IEMOCAP database is annotated by multiple annotators into categorical labels, such as anger, happiness, sadness, neutrality, as well as dimensional labels such as valence, activation and dominance.

The advantages according to the creators include, the detailed motion capture information, the interactive setting to bring out authentic emotions, and the size of the database.

3.4.2 MELD

Multimodal EmotionLines Dataset (MELD) was created by improving upon another dataset called the EmotionLines dataset by adding more modalities. It contains about 1400 dialogs with 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. It is [4]available for public use.

While both IEMOCAP and MELD seem similar to our requirement of dataset, they were experimented with, in pre-training phase in the previous version of this project and the results were not satisfactory. Owing to this, they have been ignored in this part of the project.

3.4.3 M2H2

M2H2: A Multimodal Multiparty Hindi Dataset For Humor Recognition in Conversation released by (Chauhan et al., 2021) is a dataset that was created for the task of humour recognition and is similar to our dataset in the sense that it contains conversational data. The modalities involved are text, audio and video. Instead of having dyadic conversations, the dataset contains instances that are multi-party conversations from a famous hindi series named 'Shriman Shrimathi Phir se'.

Some other details are as follows

[3] https://sail.usc.edu/iemocap/index.html

[4] https://github.com/declare-lab/MELD

- **Size:** 4.46 hours
- **Utterances:** 6191 utterances
- **Episodes:** 13 episodes
- **Language:** Hindi
- **Source:** Shriman Shrimathi Phir se

The dataset is made available publicly[5]and is organized modality-wise based on episodes into folders. Each instance is as shown in table 2

scene id	Identifies the scene
S.No	To identify the utterance within the scene
start time	Start time of the utterance
end time	End time of the utterance
utterance	The sentence uttered
label	Humour Label
speaker	Speaker information along with the utterance id to which the speaker is responding

Table 2: M2H2 Dataset Fields

This dataset could come in handy for problems related to humour or incongruity in general, particularly while working on Indian languages.

3.4.4 MaSaC

(Bedi et al., 2021) worked on the problem of **Multimodal Sarcasm detection and Humour classification**, again in the language of Hindi, but interestingly in Code-Mixed situation along with English. They publicly release their code-mixed dataset for research on github[6]. It contains both English words in Hindi and Hindi words in English as shown below in 3

Maya:blackberry के फोन्स सिग्नल्स नगरवाद करीब रखो तो सिग्नल से ब्रेन डैमेज हो सकता

Maya:abhee kisee का esemes aaega lekin मुख्य riyaakaaree तो नहीं

Figure 3: Examples from MaSaC

Similar to M2H2, MaSaC also contains humour labels for the data, however it also contains sarcasm labels for each utterance.
Some other details are as follows

- **Utterances:** 15K utterances
- **Episodes:** 50 episodes (400 scenes)
- **Language:** Hindi+English code-mixed
- **Source:** Sarabhai vs. Sarabhai

[5]https://github.com/declare-lab/M2H2-dataset

[6]https://github.com/LCS2-IIITD/MSH-COMICS.git

3.4.5 Image+Text Sarcasm Data

While looking for multimodal datasets another commonly faced situation is that, relatively more datasets labelled as multimodal consider images to be their visual modality instead of video as desired by us. Nonetheless, the following dataset created by (Sangwan et al., 2020) contains instances each of which have a text along with an image associated and was built for the task of sarcasm detection. For testing their proposed approach for sarcasm detection they compiled two Instagram based datasets.

- **Silver dataset**
  - Posts: 10K sarcastic and 10K non-sarcastic
  - Annotation method: Hashtag based
- **Gold dataset**
  - Posts: 1600 sarcastic
  - Annotation method: Manual

Since the data is based on Instagram posts, quite often the images themselves contain some text within which could also be a carrier of incongruity and hence the authors take advantage of the transcript extracted from the image and take advantage of that too by treating it as a third modality. The following figures 4 and 5 show the kind of data this dataset holds

Figure 4: Example 1: Text incongruous with image *(Caption:Someone is excited for sweater season)*

Figure 5: Example 2: Sarcasm within the transcript in the image

3.4.6 Other Multimodal Datasets

EmotionLines (Chen et al., 2018) and EmoryNLP (Zahiri and Choi, 2017) are textual datasets with conversational data, the former containing data from the TV show Friends and private Facebook messenger dialogues, while the latter was also curated from the series Friends.

3.5 Multimodal Emotion Recognition

We discuss some related work in multimodal emotion recognition.

3.5.1 Emotion Recognition in Dialogs

(Li et al., 2021) introduced an enhanced BART model using contrastive and generative loss for improving the emotion recognition in dialog data. (Ghosal et al., 2020) utilizes common-sense knowledge to enhance the utilization of context information in conversational emotion recognition.

3.5.2 MUStARD and Extended MUStARD

The paper (Chauhan et al., 2020b), explores the idea that with Sarcasm detection as primary task, the secondary tasks of emotion and sentiment classification would help improve the quality of sarcasm detection. This multi-modal multi-task framework is further explained below.
The MUStARD dataset is a multimodal video corpus with audio-visual clips from several popular TV series for research in automated sarcasm discovery. MUStARD consists of audiovisual utterances annotated with sarcasm labels. The dataset consists of 690 conversations referred to as 'samples'. Each sample is composed of multiple dialogues. The last dialog/sentence(s) is referred to as 'utterance' and is the one that is finally classified with a label, while the remaining sentences/dialogues are referred to as 'context'. As part of the work (Chauhan et al., 2020b), an extended version of MUStARD was released with additional annotation done to include implicit and explicit sentiment and emotion labels. They use an input layer followed by inter and intra modular attention in their framework to combine different modalities.
There are also works which focus only on audio as the additional modality, in addition to text or even in the absence of text like (Byun and Lee, 2021). The paper presents an extensive set of features ranging from spectral features to harmonic features and present an analysis of how these features help in the task of Speech Emotion Recognition (SER).

3.6 Multimodal Fusion

3.6.1 Novel Attention Mechanisms

(Chauhan et al., 2020a) along with MUStARD extension, also mention a novel framework with two attention mechanisms. For multi-modality to be exploited, features are to be carefully gathered from across modalities. The input representations (Text, Video with their speaker and context and audio alone with its speaker) are passed into a FFNN dense layer to get a feature vector of fixed length which is then input to the attention mechanisms.

- Inter-Segment Inter-Modal Attention: In an utterance, the relation between feature vector of segment A in one modality and vector of another segment B in a different modality is captured

  Divide to segments $\longrightarrow$ Compare different segments randomly across modalities

- Intra-Segment Inter-Modal Attention: In an utterance, the relation between same segment's feature vectors in different modalities is captured

  Concat feature vectors (TAV) $\longrightarrow$ Divide to segments $\longrightarrow$ Compare same segment across modalities

Final outputs of the two attention blocks are fed to a BiGRU and softmax for classifying the 5 labels. The tasks won't see any bias since the weights get adjusted based on all the labels.

3.6.2 Gating

(Liu et al., 2019a) presents a very detailed work which aims at ensuring good quality representation when videos are involved with other modalities like text, audio etc., Their main aim is to focus on building a compact representation which finds application in a number of video understanding tasks, such as video retrieval, clustering and summarization. To this extend they propose a multimodal fusion framework, called, 'Collaborative Gating' that ensure that video and text that correspond to each other stays similar in representation, as compared to when they are unassociated. They treat the video, audio and embedded text as three different modalities. Since this methodology internally utilizes attention, we take inspiration from this work to perform multimodal fusion in our project.

3.7 Valence-Arousal Annotation

As part of the project, instead of only going with discrete emotion annotation, we also intend to follow the continuous mode of annotation based on the dimensional model for emotions (like Plutchik's (Plutchick, 1980) and Russell's (Russell, 1980)). The following subsection discusses research papers that performed and analysed such annotation.
(Wood et al., 2018) presents a study of emotion annotation done on tweets. It mentions that among dimensional models the most common dimensions found are the following

- Valence - Evaluation-Pleasantness
- Arousal - Activation-Arousal
- Dominance - Potency-Control

more commonly referred to as the VAD model. (Preoţiuc-Pietro et al., 2016) presents work on the task of valence arousal annotation on textual data. They gather social media posts (2895) and got them annotated by two psychologically trained annotators for valence and arousal. They follow a nine-point scale for annotation and this seemed appropriate considering the degree of freedom to choose the intensity from for both dimensions. This is the reason we chose to adopt this for our own annotation task as well. The work makes interesting observations regarding the relation between the gender, age group of the annotators and the way they perceive valence, arousal and emotions in general.
(Zafeiriou et al., 2017) is another work in this direction and is quite often cited for valence arousal works in vision data. The authors had annotated 300 Youtube videos (non-enacted, hence in-the-wild). It is considered the largest database of videos studying facial affect in-the-wild. 6-8 annotators were made to annotate for valence arousal in an online mode, where the annotators gave their scores between -1 to +1 using a joystick. To be noted that the length of these video clips goes quite large with the maximum length being around 14 and a half minutes.

3.8 Data Augmentation Tools

Data Augmentation has been gaining popularity in NLP to tackle low-resource, novel problems where data is meagerly available. In this subsection we mention some of the data augmentation tools based on different methodologies.

- **AugLy:** AugLy is an open-source data augmentation library developed by Facebook which encompasses augmentation techniques from various sources and added some of its own. The way people usually transform content, is the source of motivation for AugLy's augmentation methodology. One of the major advantages of AugLy is how it supports multiple modalities. Augmentation is allowed not only in text, but also in audio and video without actually disrupting the actual content in any of them. For example, in video, we have transformations like adding blur or overlaying emoji, text etc. It can be accessed via the link.
- **MixUP:** (Feng et al., 2021) mentions the presence of several augmentation mechanisms, along with the use of MixUP as an example of Interpolation techniques for data augmentation. It is traditionally observed to be used for image augmentation and involves combining samples in different proportions to generate a new one (as the name suggests). There is also a more advanced 3D version of this introduced in the work (Chen et al., 2020).
- **TextGenie:** [7]TextGenie is also a data augmentation library that generates new samples by applying paraphrasing, mask-filling and other such NLP techniques on given data.
- **Paraphrasers:** Several off-the-shelf paraphrasers can also act as tools for data augmentation in text. Some examples include PEGASUS (Zhang et al., 2019), Sentence-BERT[8]etc.
- **Generation Models:** As mentioned in (Feng et al., 2021), fine-tuning models like GPT-2 and GPT-3 are other options available to perform data augmentation.

3.9 Eye-tracking in Sarcasm

Several works show the use of cognitive features in NLP systems. As explained before, the research community has be showing more and more interest in leveraging cognitive features, gaze features to be more specific for difficult NLP tasks like sarcasm and we discuss the same in this subsection. (Mishra

[8]https://github.com/hetpandya/textgenie
[8]https://www.sbert.net/examples/applications/paraphrase-mining/README.html

et al., 2016), (Mishra et al., 2017), (Mathias et al., 2018) mention a list of such features used in their problems.

(Mathias et al., 2020) is one survey paper that discusses the different works done using gaze features and eye-tracking in the field o NLP. Not only, does it address the difficulties faced in gathering gaze features, but also introduces basic terminologies in gaze data, different eye-tracking datasets that are available for research.

(Mishra et al., 2016) is one of the first works which had proposed a novel mechanism to enhance sarcasm detection using gaze features. The authors propose appending the textual features with gaze data gathered while reading a sarcastic statement by believing in the hypothesis that sarcasm induces distinctive reading pattern and there by eye-movement when being understood by a reader. They report a 3.7% increase in scores by using these features with statistical ML models for sarcasm detection.

(Mishra et al., 2017) points out the difficulties and challenges in manually gathering gaze data for a given dataset or a problem. The authors propose a CNN-based framework that can automatically extract features from gaze data of readers when they read the text which can then be used alongside textual features for tasks like sentiment and sarcasm detection. They highlight their choice of using CNNs based on the concept that CNNs work best with image data, being suitable to detect edges, contours etc., and that eye-movement data would also have similar patterns to be detected to extract useful cognitive features.

3.10 Ethics for Emotion Recognition

An important consideration while performing emotion recognition studies, is the ethics to be taken into account. (Mohammad, 2021) is an extensive work, that presents the various subjective problems involved while performing emotion recognition and the measures to be taken to avoid biased or problematic data or model generation.

4 Summary

In this paper, we summarized the most relevant and recent literature that was referred to in the journey of this project including some multimodal datasets in Indian languages, emotion recognition and sarcasm detection models. We also discuss fusion architectures for merging multiple modalities and present some research work around valence and arousal annotation. We discuss data augmentation tools and strategies that were studied to improve the size of the dataset and present crucial literature around the use of eye-tracking in sarcasm studies.

References

Francisca Adoma Acheampong, Chen Wenyu, and Henry Nunoo-Mensah. 2020. Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports*, 2(7):e12189.

Nourah Alswaidan and Mohamed El Bachir Menai. 2020. A survey of state-of-the-art approaches for emotion recognition in text. *Knowledge and Information Systems*, pages 1–51.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Manjot Bedi, Shivani Kumar, Md. Shad Akhtar, and Tanmoy Chakraborty. 2021. Multi-modal sarcasm detection and humor classification in code-mixed conversations. *CoRR*, abs/2105.09984.

Sung-Woo Byun and Seok-Pil Lee. 2021. A study on a speech emotion recognition system with effective acoustic features using deep learning algorithms. *Applied Sciences*, 11(4):1890.

Dushyant Singh Chauhan, SR Dhanush, Asif Ekbal, and Pushpak Bhattacharyya. 2020a. 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.

Dushyant Singh Chauhan, Dhanush S R, Asif Ekbal, and Pushpak Bhattacharyya. 2020b. 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, Online. Association for Computational Linguistics.

Dushyant Singh Chauhan, Gopendra Vikram Singh, Navonil Majumder, Amir Zadeh, Asif Ekbal, Pushpak Bhattacharyya, Louis-Philippe Morency, and Soujanya Poria. 2021. M2H2: A multimodal multiparty hindi dataset for humor recognition in conversations. *CoRR*, abs/2108.01260.

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.

Yunlu Chen, Vincent Tao Hu, Efstratios Gavves, Thomas Mensink, Pascal Mettes, Pengwan Yang, and Cees G. M. Snoek. 2020. Pointmixup: Augmentation for point clouds. *CoRR*, abs/2008.06374.

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*.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard H. Hovy. 2021. A survey of data augmentation approaches for NLP. *CoRR*, abs/2105.03075.

Deepanway Ghosal, Navonil Majumder, Alexander F. Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. COSMIC: commonsense knowledge for emotion identification in conversations. *CoRR*, abs/2010.02795.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

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

Aditya Joshi, Vaibhav Tripathi, and Pushpak Bhattacharyya. 2016b. Emotion analysis from text: A survey. *Center for Indian Language Technology Survey*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Shimin Li, Hang Yan, and Xipeng Qiu. 2021. Contrast and generation make BART a good dialogue emotion recognizer. *CoRR*, abs/2112.11202.

Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. 2019a. Use what you have: Video retrieval using representations from collaborative experts. *arXiv preprint arXiv:1907.13487*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. *CoRR*, abs/1508.04025.

Sandeep Mathias, Diptesh Kanojia, Abhijit Mishra, and Pushpak Bhattacharya. 2020. A survey on using gaze behaviour for natural language processing. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 4907–4913. International Joint Conferences on Artificial Intelligence Organization. Survey track.

Sandeep Mathias, Diptesh Kanojia, Kevin Patel, Samarth Agrawal, Abhijit Mishra, and Pushpak Bhattacharyya. 2018. Eyes are the windows to the soul: Predicting the rating of text quality using gaze behaviour. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2352–2362, Melbourne, Australia. Association for Computational Linguistics.

Abhijit Mishra, Kuntal Dey, and Pushpak Bhattacharyya. 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)*, pages 377–387, Vancouver, Canada. Association for Computational Linguistics.

Abhijit Mishra, Diptesh Kanojia, Seema Nagar, Kuntal Dey, and Pushpak Bhattacharyya. 2016. Harnessing cognitive features for sarcasm detection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1095–1104, Berlin, Germany. Association for Computational Linguistics.

Saif M. Mohammad. 2021. Ethics sheet for automatic emotion recognition and sentiment analysis. *CoRR*, abs/2109.08256.

R Plutchick. 1980. Emotion: a psychoevolutionary synthesis. *New York, Happer & Row*.

Daniel Preoţiuc-Pietro, H Andrew Schwartz, Gregory Park, Johannes Eichstaedt, Margaret Kern, Lyle Ungar, and Elisabeth Shulman. 2016. Modelling valence and arousal in facebook posts. In *Proceedings of the 7th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 9–15.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

James Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178.

Suyash Sangwan, Md Shad Akhtar, Pranati Behera, and Asif Ekbal. 2020. I didn't mean what i wrote! exploring multimodality for sarcasm detection. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Ian D. Wood, John P. McCrae, Vladimir Andryushechkin, and Paul Buitelaar. 2018. A comparison of emotion annotation approaches for text. *Information*, 9(5).

Stefanos Zafeiriou, Dimitrios Kollias, Mihalis Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. 2017. Aff-wild: Valence and arousal 'in-the-wild' challenge. pages 1980–1987.

Sayyed M. Zahiri and Jinho D. Choi. 2017. Emotion detection on TV show transcripts with sequence-based convolutional neural networks. *CoRR*, abs/1708.04299.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. *CoRR*, abs/1912.08777.