CMU-MOSEAS: A Multimodal Language Dataset for Spanish, Portuguese, German and French

Amir Zadeh[1], **Yan Sheng Cao**[2], **Simon Hessner**[1], **Paul Pu Liang**[3], **Soujanya Poria**[4], **Louis-Philippe Morency**[1]
(1) LTI, SCS, CMU, (2) SCS, CMU, (3) MLD, SCS, CMU
(4) Singapore University of Technology and Design
{abagherz,yanshenc,shessner,pliang,morency}@cs.cmu.edu
sporia@sutd.edu.sg

Access link: https://bit.ly/2Svbg9f

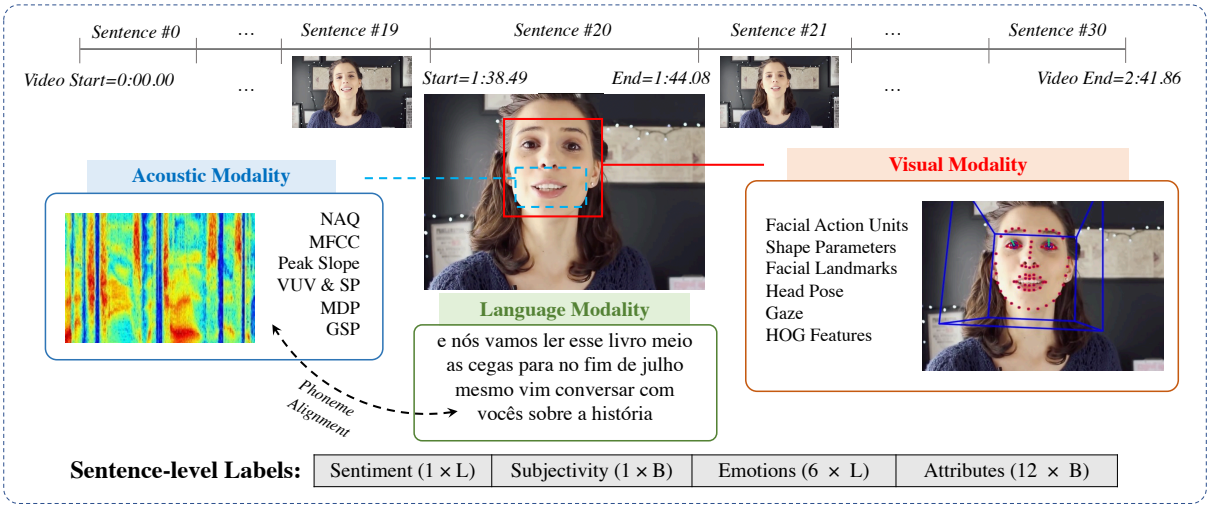

Figure 1: Overview of in-the-wild monologue videos and sentence utterances in the CMU-MOSEAS dataset. Each sentence is annotated for 20 labels including sentiment, subjectivity, emotions and attributes. "L" denotes Likert (intensity) and "B" denotes Binary for the type of the labels. The example above is a Portuguese video.

Abstract

Modeling multimodal language is a core research area in natural language processing. While languages such as English have relatively large multimodal language resources, other widely spoken languages across the globe have few or no large-scale datasets in this area. This disproportionately affects native speakers of languages other than English. As a step towards building more equitable and inclusive multimodal systems, we introduce the first large-scale multimodal language dataset for Spanish, Portuguese, German and French. The proposed dataset, called CMU-MOSEAS (CMU Multimodal Opinion Sentiment, Emotions and Attributes), is the largest of its kind with $40,000$ total labelled sentences. It covers a diverse set topics and speakers, and carries supervision of 20 labels including sentiment (and subjectivity), emotions, and attributes. Our evaluations on a state-of-the-art multimodal model demonstrates that CMU-MOSEAS enables further research for multilingual studies in multimodal language.

1 Introduction

Humans use a coordinated multimodal signal to communicate with each other. This communication signal is called multimodal language (Perniss, 2018); a complex temporal and idiosyncratic signal which includes the modalities of language, visual and acoustic. On a daily basis across the world, intentions and emotions are conveyed through joint utilization of these three modalities. While English, Chinese, and Spanish languages have resources for computational analysis of multimodal language (focusing on analysis of sentiment, subjectivity, or emotions (Yu et al., 2020; Poria et al., 2019; Zadeh et al., 2018b; Park et al., 2014; Wöllmer et al., 2013; Poria et al., 2020)), other commonly spoken languages across the globe lag behind. As Artificial Intelligence

(AI) increasingly blends into everyday life across the globe, there is a genuine need for intelligent entities capable of understanding multimodal language in different cultures. The lack of large-scale in-the-wild resources presents a substantial impediment to multilingual progress in this fundamental research area in NLP.

In this paper, we introduce a large-scale dataset for 4 languages of Spanish, Portuguese, German and French. The dataset, called CMU-MOSEAS (CMU Multimodal Opinion Sentiment, Emotions and Attributes) contains 10,000 annotated sentences from across a wide variety of speakers and topics. The dataset also contains a large subset of unlabeled samples across the 4 languages to enable unsupervised pretraining of multimodal representations. Figure 1 shows an example sentence from CMU-MOSEAS dataset along with the provided multimodal features and annotations. Annotations include sentiment, subjectivity, emotions, and attributes. We believe that data of this scale presents a step towards learning human communication at a more fine-grained level, with the long-term goal of building more equitable and inclusive NLP systems.

In the continuation of this paper, we first discuss the related resources and previous works. Subsequently, we outline the dataset creation steps, including the data acquisition, verification, and annotations. We also discuss the steps taken to protect the speakers and uphold the ethical standards of the scientific community. Finally, we experiment with a state-of-the-art multimodal language model, and demonstrate that CMU-MOSEAS presents new challenges to the NLP community.

2 Background

The related work to the content of this paper is split in two parts. We first discuss the related datasets, alongside comparisons with CMU-MOSEAS. Afterwards, we discuss the machine learning literature for modeling multimodal language.

2.1 Related Resources

We highlight the most relevant multimodal and unimodal datasets to CMU-MOSEAS. Further details of the below datasets, as well as comparison to CMU-MOSEAS is presented in Table 1.

CMU-MOSEI (Zadeh et al., 2018b) is a large-scale dataset of multimodal sentiment and emotion analysis in English. It contains over 23,000 sentences from across 1000 speakers and 250 topics. **CH-SIMS** (Yu et al., 2020) is a dataset of Chinese multimodal sentiment analysis with fine-grained annotations of sentiment per modality. **IEMOCAP** (Busso et al., 2008) is an in-lab recorded dataset which consists of 151 videos of scripted dialogues between acting participants. **POM** dataset contains 1,000 videos annotated for attributes (Park et al., 2014). The language of the dataset is English. **ICT-MMMO** (Wöllmer et al., 2013) consists of online social review videos annotated at the video level for sentiment. **CMU-MOSI** (Zadeh et al., 2016b) is a collection of 2199 opinion video clips each annotated with sentiment in the range $[-3, 3]$. **YouTube** (Morency et al., 2011) contains videos from the social media web site YouTube that span a wide range of product reviews and opinion videos. **MOUD** (Perez-Rosas et al., 2013) consists of product review videos in Spanish, annotated for sentiment. **AMMER** (Cevher et al., 2019) is a German emotion recognition dataset collected from a driver's interactions with both a virtual agent as well as a co-driver in a simulated driving environment. **UR-FUNNY** (Hasan et al., 2019) consists of more than 16000 video samples from TED talks annotated for humor. **Vera am Mittag (VAM)** (Grimm et al., 2008) corpus consists of recordings from the German TV talk-show "Vera am Mittag". This audio-visual dataset is labeled for continuous emotions of valence, activation and dominance. **RECOLA** (Ringeval et al., 2013) is an acted dataset of French language, consisting of 9.5 hours of audio, visual, and physiological (electrocardiogram, and electrodermal activity) signals. **EmoDB** (Burkhardt et al., 2005; Vondra and Vích, 2009) is a dataset of emotion recognition in German for speech and acoustic modalities.

Aside the aforementioned multimodal datasets, the following are related datasets that use only the text modality. **Stanford Sentiment Treebank (SST)** (Socher et al., 2013) includes fine grained sentiment labels for phrases in the parse trees of sentences collected from movie review data. **Large Movie Review** dataset (Maas et al., 2011) contains text from highly polar movie reviews. Textual annotated Spanish datasets have been collected from Twitter (**TASS**) (Villena Roma¡n et al., 2013-03; Pla and Hurtado, 2018; Miranda

Dataset	Samples	Speakers	Modalities	Sentiment	Emotion	Attributes	Languages	Duration
CMU-MOSEAS	40,000	1,645	l, v, a	✓	✓	✓	FR, ES, PT, DE	68 : 49
CMU-MOSEI	23,453	1000	l, v, a	✓	✓	✗	EN	65 : 53
ICT-MMMO	340	200	l, v, a	✓	✗	✗	EN	13 : 58
CMU-MOSI	2,199	98	l, v, a	✓	✗	✗	EN	02 : 36
YouTube	300	50	l, v, a	✓	✗	✗	EN	00 : 29
MOUD	400	101	l, v, a	✓	✗	✗	ES	00 : 59
IEMOCAP	10,000	10	l, v, a	✗	✓	✗	EN	11 : 28
AMMER	288	36	l, v, a	✗	✓	✗	DE	00 : 78
UR-FUNNY	16,514	1,741	l, v, a	✗	✗	✓(Humor)	EN	90 : 23
VAM	499	20	v, a	✗	✓	✗	EN	12 : 00
EmoDB	800	10	a	✗	✓	✗	DE	03 : 00
AFEW	1,645	330	v, a	✗	✓	✗	EN	02 : 28
Mimicry	48	48	v, a	✗	✓	✗	EN	11 : 00
HUMAINE	50	4	v, a	✗	✓	✗	EN	04 : 11
SEWA	538	408	v, a	✗	✓	✗	EN	04 : 39
SEMAINE	80	20	v, a	✗	✓	✗	EN	06 : 30
RECOLA	46	46	v, a	✗	✓	✗	FR	03 : 50
SST	11,855	–	l	✓	✗	✗	EN	–
Large Movie	25,000	–	l	✓	✗	✗	EN	–
TASS	3,413	–	l	✓	✗	✗	ES	–
TweetSentBR	15,000	–	l	✓	✗	✗	PT	–
SB10k	10,000	–	l	✓	✗	✗	DE	–
AM-FED	242	242	v	✗	✓	✗	EN	03 : 20

Table 1: Best viewed zoomed in. Comparison between CMU-MOSEAS and relevant datasets. CMU-MOSEAS presents a unique resource for languages of Spanish, Portuguese, German and French. $[l, v, a]$ denote [language, vision and acoustic] modalities. Duration is in the HH:MM format.

and Guzman, 2017) and hotel reviews (Molina-González et al., 2014). Polarity classification tasks based on Twitter data have also been collected in Portuguese (Brum and das Graças Volpe Nunes, 2017) (**TweetSentBR**), German (Cieliebak et al., 2017; Flender and Gips, 2017) (**SB10k**), and French (Rhouati et al., 2018). Another line of related work aims to predict humor from text in multiple languages (Castro et al., 2016, 2017).

Table 1 demonstrates that CMU-MOSEAS is a unique resource for the languages of Spanish, Portuguese, German and French.

2.2 Computational Models of Multimodal Language

Studies of multimodal language have particularly focused on the tasks of sentiment analysis (Morency et al., 2011; Yadav et al., 2015), emotion recognition (Busso et al., 2008), and personality traits recognition (Park et al., 2014). Works in this area often focus on novel multimodal neural architectures based on Transformer (Tsai et al., 2019a; Mai et al., 2019; Zadeh et al., 2019) and recurrent fusion approaches (Rahman et al., 2019; Liang et al., 2018; Zadeh et al., 2018a, 2017), as well as learning via statistical techniques such as correlation analysis (Sun et al., 2019) and tensor methods (Hou et al., 2019; Zadeh et al., 2017).

In addition to these purely discriminative approaches, recent work has also explored generative-discriminative methods for learning from multimodal language (Tsai et al., 2019b), learning from noisy or missing modalities (Mai et al., 2019; Liang et al., 2019b; Pham et al., 2019), strong baselines suitable for learning from limited data (Liang et al., 2019a), and interpretable models for language analysis (Karimi, 2018; Zadeh et al., 2018b). Several other lines of work have focuses on building stronger unimodal representations such as language (Kordjamshidi et al., 2017; Beinborn et al., 2018) and speech (Sanabria et al., 2018; Lakomkin et al., 2019; Gu et al., 2019) for multimodal language understanding.

3 CMU-MOSEAS (CMU Multimodal Opinion Sentiment, Emotions and Attributes) Dataset

The CMU-MOSEAS dataset covers 4 languages of Spanish (>500M total speakers globally), Portuguese (>200M speakers globally), German (>200M speakers globally), and French (>200M speakers globally). These languages either have Romance or Germanic roots (Renfrew, 1989). They originate from Europe, which is also the main region for our video acquisition. The languages are also spoken in the American continent

Language	Spanish	Portuguese	German	French
Total number of videos	1,000	1,000	1,000	1,000
Total number of sentences	29,544	34,633	30,549	34,042
Total number of annotated sentences	10,000	10,000	10,000	10,000
Total number of distinct speakers	341	399	480	425
Total number of distinct topics	250	250	250	250
Average length of videos (in sentences)	29.54	34.63	30.55	34.04
Average length of videos (in words)	582.22	606.23	361.00	646.84
Average length of videos (in seconds)	210.95	217.29	218.79	208.66
Average length of sentences (in words)	17.67	16.72	13.13	16.63
Average length of sentences (in seconds)	6.70	5.77	6.70	5.63
Speech rate (number of words per second)	2.76	2.79	1.65	3.10
Vocabulary size	36,120	34,982	37,969	41,762

Table 2: CMU-MOSEAS multimedia and linguistic statistics for languages of Spanish, Portuguese, German and French.

(north and south), as well as portions of Africa and the Caribbean (with different dialects, however, the European dialect is mostly comprehensible across different regions with some exceptions[1]).

Subsequently, in this section, we discuss the data acquisition and verification process, followed by outlining the annotated labels. We prioritize important details in the body of the main paper, and refer the reader to supplementary material for extra details about the dataset.

3.1 Acquisition and Verification

Monologue videos offer a rich source of multimodal language across different identities, genders, and topics. Users share their opinions online on a daily basis on websites such as YouTube[2]. In this paper, the process of finding and manually verifying monologue videos falls into the following 3 main steps:

Monologue Acquisition: In this step, monologue videos are manually found from across YouTube, using a diverse set of more than 250 search terms (see supplementary for search terms). The following regions are chosen for each language: [Spanish: Spain], [Portuguese: Portugal], [German: Germany and Austria], [French: France]. The YouTube search parameters are set based on the correct language and region. No more than 5 videos are gathered from individual channels to ensure diversity across speakers (average video to speaker ratio is 2.43 across the dataset). Only monologues with high video and audio quality are acquired. A particular focus in this step has been to acquire a set of gender-balanced videos for each language and region.

Monologue Verification: The acquired monologues in the previous step are subsequently checked by 2 native speakers of each language to ensure: 1) the language is correct and understandable, 2) the region is correct, 3) gathered transcription is high-quality, 4) the grammar and punctuation in transcriptions are correct (the transcripts are also corrected for errors). Only the videos that passed all the filters are allowed to pass this step.

Forced Alignment Verification: The text-audio synchronization is an essential step for in-depth studies of multimodal language. It allows for modeling intermodal relations at the word or phoneme levels using continuous alignment (Chen et al., 2017). All the languages in CMU-MOSEAS have pre-trained acoustic and G2P (Grapheme-2-Phoneme) models which allow for forced alignment between text and audio. The monologue videos are subsequently aligned using MFA - Montreal Forced Aligner (McAuliffe et al., 2017). Afterwards, the forced alignment output is manually checked by native speakers to ensure the high quality of the alignment.

Utilizing the above pipeline, a total of 1,000 monologue videos for each language of Spanish, Portuguese, German, and French are acquired (over the course of two years). From across these

[1] Such as Swiss German.

[2] With licenses allowing for fair usage of their content https://www.youtube.com/intl/en-GB/about/copyright/fair-use/

videos, a total of $10,000$ sentences are annotated according to Section 3.3. The sentence splitting follows a similar procedure as reported in the creation of CMU-MOSEI. Therefore, the size of the dataset is a total of $40,000$ annotated samples ($10,000$ for each language), accompanied by a large unsupervised set of sentences for each language. Table 2 shows the overall statistics of the data (see Section 3.6 for the methodology of face identification).

3.2 Privacy and Ethics

A specific focus of CMU-MOSEAS is on protecting the privacy of the speakers. Even though videos are publicly available on YouTube, a specific EULA (End User License Agreement) is required to download the labels (to see the EULA, please refer to supplementary). Non-invertible high-level computational features are provided publicly online. These features cannot be inverted to recreate the video or audio. For example, FAU (Facial Action Units) intensities. In simple terms, no speaker can deterministically be identified by these features.

3.3 Annotator Selection

Annotation of videos in CMU-MOSEAS is done by crowd workers[3] of the Amazon Mechanical Turk (AMT) platform. The workers are filtered to have higher than 95% acceptance rate over at least $5,000$ completed jobs. The annotators are native speakers of the languages discussed in Section 3.1. For each annotation, the annotators are given a sentence utterance and asked to annotate the labels of CMU-MOSEAS (discussed in Section 3.4). Labels are arranged on a web-page which allows the users to annotate them after watching the sentence utterance. At the beginning of the annotation process, the annotators are given a 5 minute training video describing the annotation scheme in their respective language (see supplementary for annotation user interface and training material). Each sentence utterance is annotated by 3 distinct annotators. Annotations are subsequently checked for criteria such as the speed of annotation, or answering secret key questions. Annotators with poor performance are subsequently removed.

[3] AMT screens annotators and tags reliable ones as Master workers.

3.4 Labels

The labels and an overview of their annotation scheme is as follows. Labels are annotated based on Likert (i.e. intensity) or Binary steps. Labels are checked via cyclic translation to eliminate divergence in their meaning caused by language barriers. Annotation scheme also help in this regard since all languages follow the same translation method, closely supervised by the authors of this paper.

Sentiment (Likert): We follow a similar annotation scheme as designed in prior literature for multimodal sentiment analysis (Morency et al., 2011; Wöllmer et al., 2013), and closely inspired by utterance sentiment annotations (Zadeh et al., 2016b). Sentence utterances are individually annotated for their perceived sentiment (i.e. the sentiment of the speaker in the video). Each sentence is annotated for sentiment on a $[-3, 3]$ Likert scale of: $[-3$: highly negative, -2 negative, -1 weakly negative, 0 neutral, $+1$ weakly positive, $+2$ positive, $+3$ highly positive$]$.

Subjectivity (Binary): The sentence utterances are annotated for whether or not the speaker expresses an opinion, as opposed to a factual statement (Wiebe et al., 2005). Subjectivity can be conveyed through either an explicit or implicit mention of a private state (Zadeh et al., 2016a), both of which are included in the annotation scheme.

Emotions (Likert): Ekman emotions (Ekman et al., 1980) of $\{$Happiness (HA), Sadness (SA), Anger (AN), Fear (FE), Disgust (DI), Surprise (SU)$\}$ are annotated on a $[0, 3]$ Likert scale for presence of emotion x: $[0$: no evidence of x, 1: weakly x, 2: x, 3: highly $x]$. Parenthesis denotes column name in Table 3. Sentence utterances are annotated for their perceived emotions (speaker's emotions).

Attributes (Binary): The attribute annotations are inspired by Persuasion Opinion Multimodal (POM) Dataset (Park et al., 2014) and follows a similar annotation scheme. The annotators are asked for their opinion about certain attributes being applicable to the speaker or the utterance (sentence). The following attributes are annotated: Dominant (DO), Confident (CO), Passion-

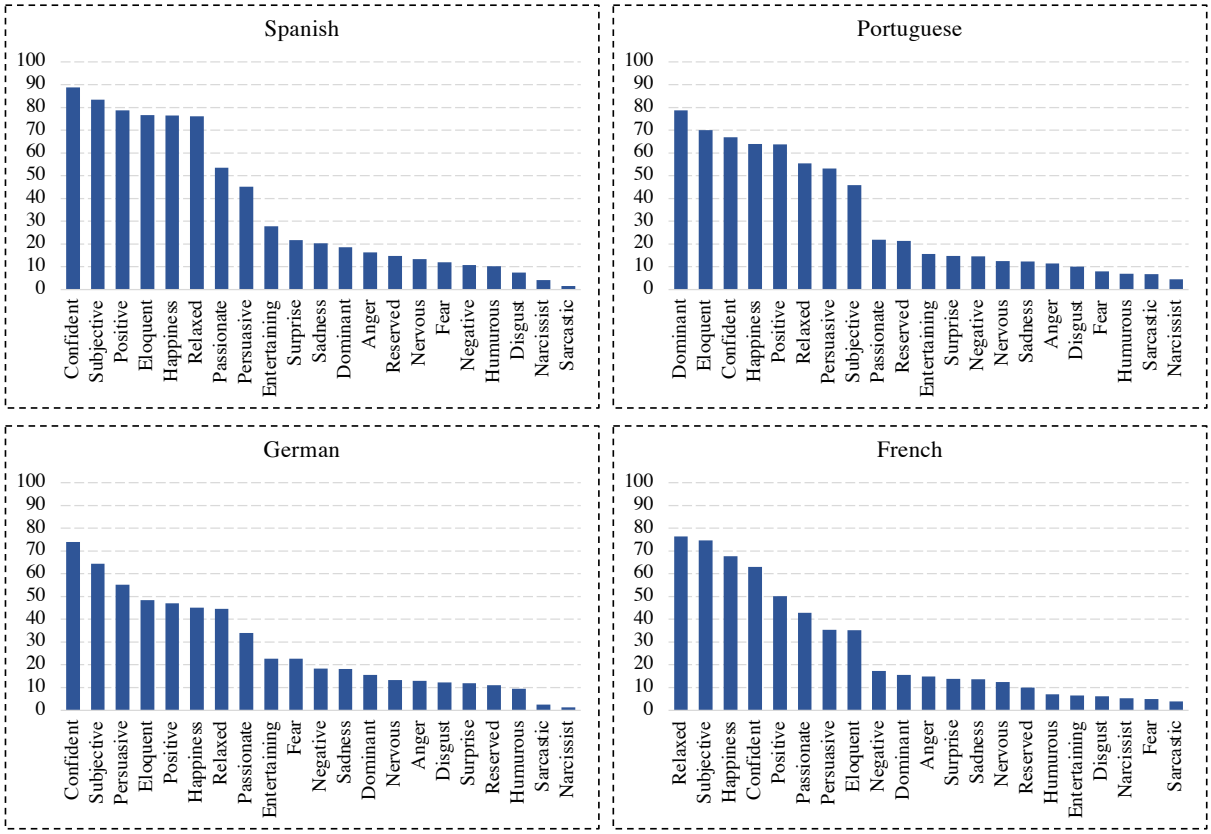

Figure 2: Label statistics of the CMU-MOSEAS. y-axis denotes the percentage of the label being present, and x-axis denotes the sentiment, subjectivity, emotions, and personality attribute labels. "Positive" and "Negative" denote sentiment.

ate (PA), Persuasive (PE), Relaxed (REL), Eloquent (EL), Nervous (NE), Entertaining (EN), Reserved (RES), Narcissist (NA), Sarcastic (SAR), and Humorous (HU). Similar to emotions, parenthesis denotes the column in Table 3.

3.5 Label Statistics

A unique aspect of CMU-MOSEAS is allowing for multimodal statistical comparisons between various languages. We outline some preliminary such comparisons in this section. Figure 2 shows the distribution of labels for CMU-MOSEAS dataset. The individual labels across different languages roughly follow a similar distribution. However, subtle differences exemplify a unique characteristic for each language.

The data suggests that perception of dominance in Portuguese may be fundamentally different than other languages. While dominance is neither a sparse nor a common label for Spanish, German and French, in Portuguese it is the most common label.

Positive sentiment seems to be reported more commonly in Spanish videos. German and French report a near toss-up for positive as opposed to negative or neutral combined (non-positive). Note, English also follows near toss-up between the positive vs non-positive (Zadeh et al., 2018b). Spanish and Portuguese report positive sentiment more commonly.

Spanish videos are more commonly labelled as confident than other languages, while other languages are at a similar level for this label.

Perception of relaxed attribute is also different across languages. French subset reports the relaxed label as the most common among labels. Overall French and Spanish are higher in this attribute than German and Portuguese.

Positive and Happiness labels closely follow each other, except for French language.

A large portion of the sentences are subjective as they convey personal opinions (as opposed to factual statements such as news broadcast).

Labels such as sadness, anger, humorous, and narcissist are similarly distributed between the languages.

Majority of labels have at least $1,000$ data points. Some labels are less frequent than others. This is aligned with findings from previous datasets for emotions (Zadeh et al., 2018b) and attributes (Park et al., 2014). For example, sarcasm is a rare attribute, even in entertainment and comedy TV shows (Castro et al., 2019).

Overall, languages seem to have intriguing similarities and differences which CMU-MOSEAS allows for studying.

3.6 Multimodal Feature Extraction

Data points in CMU-MOSEAS come in video format and include three main modalities. The extracted descriptors for each modality are as follows:

Language and Forced Alignment: All videos in CMU-MOSEAS have manual and punctuated transcriptions. Transcriptions are checked and corrected for both (see Section 3.1). Punctuation markers are used to separate sentences, similar to CMU-MOSEI. Words and audio are aligned at phoneme level using Montreal Forced Aligner (McAuliffe et al., 2017). This alignment is subsequently manually checked and corrected.

Visual: Frames are extracted from the full videos at 30Hz. The bounding box of the face is extracted using the RetinaFace (Deng et al., 2019b). Identities are extracted using ArcFace (Deng et al., 2019a). The parameters of both tools are tuned to reflect the correct number of identities. MultiComp OpenFace 2.0 (Baltrusaitis et al., 2018) is used to extract facial action units (depicting facial muscle movements), facial shape parameters (acquired using a projected latent shape by Structure from Motion), facial landmarks (68 3D landmarks on inside and boundary of face), head pose (position and Euler angles) and eye gaze (Euler angles). Visual feature extraction is done at $30Hz$.

Acoustic: We use the COVAREP software (Degottex et al., 2014) to extract acoustic features including 12 Mel-frequency cepstral coefficients, pitch tracking and voiced/unvoiced segmenting features (Drugman and Alwan, 2011), glottal source parameters (Childers and Lee, 1991; Drugman et al., 2012; Titze and Sundberg, 1992; Alku, 1992; Alku et al., 1997, 2002), peak slope parameters and maxima dispersion quotients (Kane and Gobl, 2013). Similar features are also extracted using OpenSmile (Eyben et al., 2010). Acoustic feature extraction is done at $100Hz$.

Dataset and features are available for download from the CMU Multimodal SDK, via the link `https://bit.ly/2Svbg9f`. This link provides the most accurate and up to date scoreboard, features and announcements for future readers. The original videos require submission of an EULA to the authors of this paper. EULA may change to reflect the latest privacy rules. Users in different countries and jurisdictions may need to submit additional forms.

4 Experimental Baselines

In this section we establish baselines for CMU-MOSEAS dataset. We choose a state of the art transformer-based neural model for this purpose. The model has shown state-of-the-art performance across several multimodal language tasks including multimodal sentiment analysis and emotion recognition. The CMU-MOSEAS dataset is split in the folds of train, validation and test (available on the CMU Multimodal SDK). What follows is a brief description of the baseline model.

Multimodal Transformer (MulT): Multimodal Transformer (Tsai et al., 2019a) is an extension of the well-known Transformer model (Vaswani et al., 2017) to multimodal time-series data. Each modality has a separate Transformer encoding the information hierarchically. The key component of MulT is a set of cross-modal attention blocks that cross-attend between time-series data from two modalities. MulT is among state-of-the-art models on both aligned and unaligned versions of the CMU-MOSEI and CMU-MOSI datasets. We use the author provided code for these experiments[4], with learning rate of $10e-4$ and the Adam optimizer (Kingma and Ba, 2014). The Transformer hidden unit size is 40 with 4 cross-modal blocks and 10 attention heads. Dropout is universally set at 0.1. The best model is chosen using the validation set of each language. We use the aligned variant of MulT.

For each language, we perform word-level alignment to acquire the expectation of visual

[4]https://github.com/yaohungt/Multimodal-Transformer

Task	Model	Sent.	Subj.	Emotions						Attributes											
				HA	SA	AN	DI	SU	FE	DO	CO	PA	PE	REL	EL	NE	EN	RES	NA	HU	SAR
ES	MulT	0.59	0.71	0.58	0.66	0.57	0.64	0.68	0.69	0.74	0.67	0.57	0.56	0.66	0.69	0.71	0.64	0.69	0.79	0.70	0.79
DE	MulT	0.64	0.74	0.63	0.70	0.69	0.73	0.70	0.72	0.66	0.59	0.62	0.51	0.62	0.68	0.69	0.74	0.61	0.71	0.79	0.74
FR	MulT	0.60	0.68	0.60	0.73	0.63	0.68	0.59	0.64	0.69	0.65	0.57	0.50	0.52	0.59	0.65	0.67	0.70	0.84	0.77	0.81
PT	MulT	0.62	0.70	0.59	0.62	0.68	0.77	0.66	0.76	0.59	0.60	0.61	0.56	0.58	0.61	0.73	0.71	0.64	0.80	0.72	0.76

Table 3: Results of baseline experiments on CMU-MOSEAS dataset, using MulT neural model. The reported measure is weighted F1 score. Results indicate that the current state of the art is still far from desirable performance. A special focus enabled by CMU-MOSEAS is generalization of such models to multilingual scenarios.

and acoustic contexts per word (Chen et al., 2017), identical to the methodology used by MulT (aligned variant). The maximum sequence length is set at 50. Sequences are padded on the left with zeros. For language, we use the one-hot representation of the words. For acoustic, we concatenate COVAREP and OpenSmile features. The experiments are performed tri-label for sentiment (negative, neutral, positive) and binary for emotions and attributes; similar methodology is employed by MulT. The above models are trained to minimize Mean-Absolute Error (MAE). The metric used to evaluate model performance is the F1 measure, which is a more suitable metric when there are imbalanced classes as is the case for some labels in our dataset (i.e. rare attributes). For extra details of experiments, as well as other results including MAE and correlation, please refer to the github.

Table 3 reports the F1 measure for the performance of MulT over different languages in the CMU-MOSEAS dataset. Information from all modalities are used as input to the model. While the model is capable of predicting the labels from multimodal data to some extent, the performance is still far from perfect. Therefore, we believe the CMU-MOSEAS dataset brings new challenges to the field of NLP and modeling multimodal language.

5 Conclusion

In this paper, we introduced a new large-scale in-the-wild dataset of multimodal language, called CMU-MOSEAS (CMU Multimodal Opinion Sentiment, Emotions and Attributes). The CMU-MOSEAS dataset is the largest of its kind in all four constituent languages (French, German, Portuguese, and Spanish) with 40,000 total samples spanning 1,645 speakers and 250 topics. CMU-MOSEAS contains 20 annotated labels including sentiment (and subjectivity), emotions, and personality traits. The dataset and accompanied descriptors will be made publicly available, and regularly updated with new feature descriptors as multimodal learning advances. To protect the privacy of the speakers, the released descriptors will not carry invertible information, and no video or audio can be reconstructed based on the extracted features. A state-of-the-art model was trained to establish strong baselines for future studies. We believe that data of this scale presents a step towards learning human communication at a more fine-grained level, with the long-term goal of building more equitable and inclusive NLP systems across multiple languages.

Acknowledgement

This material is based upon work partially supported by the National Science Foundation (Awards #1750439, #1722822, #1734868) and National Institutes of Health. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of National Science Foundation or National Institutes of Health, and no official endorsement should be inferred. We also would like to thank those students at Carnegie Mellon University who contributed to this project in various ways, including those who contributed to manual acquisition and verification of videos or transcripts.

References

Paavo Alku. 1992. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech communication*, 11(2-3):109–118.

Paavo Alku, Tom Bäckström, and Erkki Vilkman. 2002. Normalized amplitude quotient for parametrization of the glottal flow. *the Journal of the Acoustical Society of America*, 112(2):701–710.

Paavo Alku, Helmer Strik, and Erkki Vilkman. 1997. Parabolic spectral parameter—a new method for quantification of the glottal flow. *Speech Communication*, 22(1):67–79.

Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66. IEEE.

Lisa Beinborn, Teresa Botschen, and Iryna Gurevych. 2018. Multimodal grounding for language processing. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2325–2339, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Henrico Bertini Brum and Maria das Graças Volpe Nunes. 2017. Building a sentiment corpus of tweets in brazilian portuguese. *CoRR*, abs/1712.08917.

Felix Burkhardt, Astrid Paeschke, M. Rolfes, Walter F. Sendlmeier, and Benjamin Weiss. 2005. A database of german emotional speech. In *INTERSPEECH*, pages 1517–1520. ISCA.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Journal of Language Resources and Evaluation*, 42(4):335–359.

Santiago Castro, Matías Cubero, Diego Garat, and Guillermo Moncecchi. 2016. Is this a joke? detecting humor in spanish tweets. In *Advances in Artificial Intelligence - IBERAMIA 2016*, pages 139–150, Cham. Springer International Publishing.

Santiago Castro, Matías Cubero, Diego Garat, and Guillermo Moncecchi. 2017. HUMOR: A crowd-annotated spanish corpus for humor analysis. *CoRR*, abs/1710.00477.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards multimodal sarcasm detection (an _obviously_ perfect paper). *arXiv preprint arXiv:1906.01815*.

Deniz Cevher, Sebastian Zepf, and Roman Klinger. 2019. Towards multimodal emotion recognition in german speech events in cars using transfer learning. In *Proceedings of the 15th Conference on Natural Language Processing, KONVENS 2019, Erlangen, Germany, October 9-11, 2019*.

Minghai Chen, Sen Wang, Paul Pu Liang, Tadas Baltrušaitis, Amir Zadeh, and Louis-Philippe Morency. 2017. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 163–171. ACM.

Donald G Childers and CK Lee. 1991. Vocal quality factors: Analysis, synthesis, and perception. *the Journal of the Acoustical Society of America*, 90(5):2394–2410.

Mark Cieliebak, Jan Milan Deriu, Dominic Egger, and Fatih Uzdilli. 2017. A twitter corpus and benchmark resources for German sentiment analysis. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 45–51, Valencia, Spain. Association for Computational Linguistics.

Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. Covarep—a collaborative voice analysis repository for speech technologies. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 960–964. IEEE.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019a. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699.

Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. 2019b. Retinaface: Single-stage dense face localisation in the wild. *arXiv preprint arXiv:1905.00641*.

Thomas Drugman and Abeer Alwan. 2011. Joint robust voicing detection and pitch estimation based on residual harmonics. In *Interspeech*, pages 1973–1976.

Thomas Drugman, Mark Thomas, Jon Gudnason, Patrick Naylor, and Thierry Dutoit. 2012. Detection of glottal closure instants from speech signals: A quantitative review. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(3):994–1006.

Paul Ekman, Wallace V Freisen, and Sonia Ancoli. 1980. Facial signs of emotional experience. *Journal of personality and social psychology*, 39(6):1125.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462. ACM.

Malte Flender and Carsten Gips. 2017. Sentiment analysis of a german twitter-corpus. In *LWDA*.

Michael Grimm, Kristian Kroschel, and Shrikanth Narayanan. 2008. The vera am mittag german audio-visual emotional speech database. In *ICME*, pages 865–868. IEEE.

Yue Gu, Xinyu Lyu, Weijia Sun, Weitian Li, Shuhong Chen, Xinyu Li, and Ivan Marsic. 2019. Mutual correlation attentive factors in dyadic fusion networks for speech emotion recognition. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, pages 157–166, New York, NY, USA. ACM.

Md Kamrul Hasan, Wasifur Rahman, Amir Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, et al. 2019. Ur-funny: A multimodal language dataset for understanding humor. *arXiv preprint arXiv:1904.06618*.

Ming Hou, Jiajia Tang, Jianhai Zhang, Wanzeng Kong, and Qibin Zhao. 2019. Deep multimodal multilinear fusion with high-order polynomial pooling. In *Advances in Neural Information Processing Systems*, pages 12113–12122.

John Kane and Christer Gobl. 2013. Wavelet maxima dispersion for breathy to tense voice discrimination. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(6):1170–1179.

Hamid Karimi. 2018. Interpretable multimodal deception detection in videos. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, ICMI '18, pages 511–515, New York, NY, USA. ACM.

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

Parisa Kordjamshidi, Taher Rahgooy, and Umar Manzoor. 2017. Spatial language understanding with multimodal graphs using declarative learning based programming. In *Proceedings of the 2nd Workshop on Structured Prediction for Natural Language Processing*, pages 33–43, Copenhagen, Denmark. Association for Computational Linguistics.

Egor Lakomkin, Mohammad-Ali Zamani, Cornelius Weber, Sven Magg, and Stefan Wermter. 2019. Incorporating end-to-end speech recognition models for sentiment analysis. *CoRR*, abs/1902.11245.

Paul Pu Liang, Yao Chong Lim, Yao-Hung Hubert Tsai, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2019a. Strong and simple baselines for multimodal utterance embeddings. *CoRR*, abs/1906.02125.

Paul Pu Liang, Zhun Liu, Yao-Hung Hubert Tsai, Qibin Zhao, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2019b. Learning representations from imperfect time series data via tensor rank regularization. *CoRR*, abs/1907.01011.

Paul Pu Liang, Ziyin Liu, Amir Zadeh, and Louis-Philippe Morency. 2018. Multimodal language analysis with recurrent multistage fusion. *EMNLP*.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

S. Mai, S. Xing, and H. Hu. 2019. Locally confined modality fusion network with a global perspective for multimodal human affective computing. *IEEE Transactions on Multimedia*, pages 1–1.

Sijie Mai, Haifeng Hu, and Songlong Xing. 2019. Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, pages 498–502.

Carlos Henriquez Miranda and Jaime Guzman. 2017. A Review of Sentiment Analysis in Spanish. *Tecciencia*, 12:35 – 48.

M. Dolores Molina-González, Eugenio Martínez-Cámara, María Teresa Martín-Valdivia, and Luis Alfonso Ureña López. 2014. Cross-domain sentiment analysis using spanish opinionated words. In *NLDB*.

Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. 2011. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th International Conference on Multimodal Interactions*, pages 169–176. ACM.

Sunghyun Park, Han Suk Shim, Moitreya Chatterjee, Kenji Sagae, and Louis-Philippe Morency. 2014. Computational analysis of persuasiveness in social multimedia: A novel dataset and multimodal prediction approach. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 50–57. ACM.

Veronica Perez-Rosas, Rada Mihalcea, and Louis-Philippe Morency. 2013. Utterance-Level Multimodal Sentiment Analysis. In *Association for Computational Linguistics (ACL)*, Sofia, Bulgaria.

Pamela Perniss. 2018. Why we should study multimodal language. *Frontiers in psychology*, 9:1109.

Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabas Poczos. 2019. Found in translation: Learning robust joint representations by cyclic translations between modalities. *AAAI*.

Ferran Pla and Lluís-F. Hurtado. 2018. Spanish sentiment analysis in twitter at the tass workshop. *Lang. Resour. Eval.*, 52(2):645–672.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, and Rada Mihalcea. 2020. Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research. *arXiv preprint arXiv:2005.00357*.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536.

Wasifur Rahman, Md Kamrul Hasan, Amir Zadeh, Louis-Philippe Morency, and Mohammed Ehsan Hoque. 2019. M-bert: Injecting multimodal information in the bert structure. *arXiv preprint arXiv:1908.05787*.

Colin Renfrew. 1989. The origins of indo-european languages. *Scientific American*, 261(4):106–115.

Abdelkader Rhouati, Jamal Berrich, Mohammed Belkasmi, and Bouchentouf Toumi. 2018. Sentiment analysis of french tweets based on subjective lexicon approach: Evaluation of the use of opennlp and corenlp tools. *Journal of Computer Science*, 14:829–836.

Fabien Ringeval, Andreas Sonderegger, Jürgen S. Sauer, and Denis Lalanne. 2013. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *FG*, pages 1–8. IEEE Computer Society.

Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: A large-scale dataset for multimodal language understanding. *CoRR*, abs/1811.00347.

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, Christopher Potts, et al. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, volume 1631, page 1642. Citeseer.

Zhongkai Sun, Prathusha K. Sarma, William A. Sethares, and Yingyu Liang. 2019. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. *CoRR*, abs/1911.05544.

Ingo R Titze and Johan Sundberg. 1992. Vocal intensity in speakers and singers. *the Journal of the Acoustical Society of America*, 91(5):2936–2946.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019a. Multimodal transformer for unaligned multimodal language sequences. *arXiv preprint arXiv:1906.00295*.

Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019b. Learning factorized multimodal representations. *ICLR*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Julio Villena Roma¡n, Sara Lana Serrano, Eugenio MartÃnez CÃ¡mara, and JosÃ© Carlos GonzÃ¡lez CristÃ³bal. 2013-03. Tass - workshop on sentiment analysis at sepln.

Martin Vondra and Robert Vích. 2009. Recognition of emotions in german speech using gaussian mixture models. In *Multimodal Signals: Cognitive and Algorithmic Issues*, pages 256–263, Berlin, Heidelberg. Springer Berlin Heidelberg.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3):165–210.

Martin Wöllmer, Felix Weninger, Tobias Knaup, Björn Schuller, Congkai Sun, Kenji Sagae, and Louis-Philippe Morency. 2013. Youtube movie reviews: Sentiment analysis in an audio-visual context. *IEEE Intelligent Systems*, 28(3):46–53.

S. K. Yadav, M. Bhushan, and S. Gupta. 2015. Multimodal sentiment analysis: Sentiment analysis using audiovisual format. In *2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 1415–1419.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. 2020. Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3718–3727.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Empirical Methods in Natural Language Processing, EMNLP*.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018a. Memory fusion network for multi-view sequential learning. *arXiv preprint arXiv:1802.00927*.

Amir Zadeh, Chengfeng Mao, Kelly Shi, Yiwei Zhang, Paul Pu Liang, Soujanya Poria, and Louis-Philippe Morency. 2019. Factorized multimodal transformer for multimodal sequential learning. *arXiv preprint arXiv:1911.09826*.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016a. Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*.

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

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018b. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *ACL*.