Context-Dependent Sentiment Analysis in User-Generated Videos

Soujanya Poria
Temasek Laboratories
NTU, Singapore
sporia@ntu.edu.sg

Erik Cambria
School of Computer Science and Engineering, NTU, Singapore
cambria@ntu.edu.sg

Devamanyu Hazarika
Computer Science and Engineering, NITW, India
devamanyu@sentic.net

Navonil Mazumder
Centro de Investigacin en Computacin, IPN, Mexico
navonil@sentic.net

Amir Zadeh
Language Technologies Institute, CMU, USA
abagherz@cs.cmu.edu

Louis-Philippe Morency
Language Technologies Institute, CMU, USA
morency@cs.cmu.edu

Abstract

Multimodal sentiment analysis is a developing area of research, which involves the identification of sentiments in videos. Current research considers utterances as independent entities, i.e., ignores the interdependencies and relations among the utterances of a video. In this paper, we propose a LSTM-based model that enables utterances to capture contextual information from their surroundings in the same video, thus aiding the classification process. Our method shows 5-10% performance improvement over the state of the art and high robustness to generalizability.

1 Introduction

Sentiment analysis is a 'suitcase' research problem that requires tackling many NLP sub-tasks, e.g., aspect extraction (Poria et al., 2016a), named entity recognition (Ma et al., 2016), concept extraction (Rajagopal et al., 2013), sarcasm detection (Poria et al., 2016b), personality recognition (Majumder et al., 2017), and more.

Sentiment analysis can be performed at different granularity levels, e.g., subjectivity detection simply classifies data as either subjective (opinionated) or objective (neutral), while polarity detection focuses on determining whether subjective data indicate positive or negative sentiment. Emotion recognition further breaks down the inferred polarity into a set of emotions conveyed by the subjective data, e.g., positive sentiment can be caused by joy or anticipation, while negative sentiment can be caused by fear or disgust.

Even though the primary focus of this paper is to classify sentiment in videos, we also show the performance of the proposed method for the finer-grained task of emotion recognition.

Emotion recognition and sentiment analysis have become a new trend in social media, helping users and companies to automatically extract the opinions expressed in user-generated content, especially videos. Thanks to the high availability of computers and smartphones, and the rapid rise of social media, consumers tend to record their reviews and opinions about products or films and upload them on social media platforms, such as YouTube and Facebook. Such videos often contain comparisons, which can aid prospective buyers make an informed decision.

The primary advantage of analyzing videos over text is the surplus of behavioral cues present in vocal and visual modalities. The vocal modulations and facial expressions in the visual data, along with textual data, provide important cues to better identify affective states of the opinion holder. Thus, a combination of text and video data helps to create a more robust emotion and sentiment analysis model (Poria et al., 2017a).

An utterance (Olson, 1977) is a unit of speech bound by breathes or pauses. Utterance-level sentiment analysis focuses on tagging every utterance of a video with a sentiment label (instead of assigning a unique label to the whole video). In particular, utterance-level sentiment analysis is useful to understand the sentiment dynamics of different aspects of the topics covered by the speaker throughout his/her speech.

Recently, a number of approaches to multimodal sentiment analysis, producing interesting results, have been proposed (Pérez-Rosas et al., 2013; Wollmer et al., 2013; Poria et al., 2015). However, there are major issues that remain unaddressed. Not considering the relation and dependencies among the utterances is one of such issues. State-of-the-art approaches in this area treat utterances independently and ignore the order of utterances in a video (Cambria et al., 2017b).

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pages 873–883
Vancouver, Canada, July 30 - August 4, 2017.
https://doi.org/10.18653/v1/P17-1081

Every utterance in a video is spoken at a distinct time and in a particular order. Thus, a video can be treated as a sequence of utterances. Like any other sequence classification problem (Collobert et al., 2011), sequential utterances of a video may largely be contextually correlated and, hence, influence each other's sentiment distribution. In our paper, we give importance to the order in which utterances appear in a video.

We treat surrounding utterances as the context of the utterance that is aimed to be classified. For example, the MOSI dataset (Zadeh et al., 2016) contains a video, in which a girl reviews the movie 'Green Hornet'. At one point, she says "The Green Hornet did something similar". Normally, doing something similar, i.e., monotonous or repetitive might be perceived as negative. However, the nearby utterances "It engages the audience more", "they took a new spin on it", "and I just loved it" indicate a positive context.

The hypothesis of the independence of tokens is quite popular in information retrieval and data mining, e.g., bag-of-words model, but it has a lot limitations (Cambria and White, 2014). In this paper, we discard such an oversimplifying hypothesis and develop a framework based on long short-term memory (LSTM) that takes a sequence of utterances as input and extracts contextual utterance-level features.

The other uncovered major issues in the literature are the role of speaker-dependent versus speaker-independent models, the impact of each modality across the dataset, and generalization ability of a multimodal sentiment classifier. Leaving these issues unaddressed has presented difficulties in effective comparison of different multimodal sentiment analysis methods. In this work, we address all of these issues.

Our model preserves the sequential order of utterances and enables consecutive utterances to share information, thus providing contextual information to the utterance-level sentiment classification process. Experimental results show that the proposed framework has outperformed the state of the art on three benchmark datasets by 5-10%.

The paper is organized as follows: Section 2 provides a brief literature review on multimodal sentiment analysis; Section 3 describes the proposed method in detail; experimental results and discussion are shown in Section 4; finally, Section 5 concludes the paper.

2 Related Work

The opportunity to capture people's opinions has raised growing interest both within the scientific community, for the new research challenges, and in the business world, due to the remarkable benefits to be had from financial market prediction.

Text-based sentiment analysis systems can be broadly categorized into knowledge-based and statistics-based approaches (Cambria et al., 2017a). While the use of knowledge bases was initially more popular for the identification of polarity in text (Cambria et al., 2016; Poria et al., 2016c), sentiment analysis researchers have recently been using statistics-based approaches, with a special focus on supervised statistical methods (Socher et al., 2013; Oneto et al., 2016).

In 1974, Ekman (Ekman, 1974) carried out extensive studies on facial expressions which showed that universal facial expressions are able to provide sufficient clues to detect emotions. Recent studies on speech-based emotion analysis (Datcu and Rothkrantz, 2008) have focused on identifying relevant acoustic features, such as fundamental frequency (pitch), intensity of utterance, bandwidth, and duration.

As for fusing audio and visual modalities for emotion recognition, two of the early works were (De Silva et al., 1997) and (Chen et al., 1998). Both works showed that a bimodal system yielded a higher accuracy than any unimodal system. More recent research on audio-visual fusion for emotion recognition has been conducted at either feature level (Kessous et al., 2010) or decision level (Schuller, 2011). While there are many research papers on audio-visual fusion for emotion recognition, only a few have been devoted to multimodal emotion or sentiment analysis using textual clues along with visual and audio modalities. (Wollmer et al., 2013) and (Rozgic et al., 2012) fused information from audio, visual, and textual modalities to extract emotion and sentiment.

Poria et al. (Poria et al., 2015, 2016d, 2017b) extracted audio, visual and textual features using convolutional neural network (CNN); concatenated those features and employed multiple kernel learning (MKL) for final sentiment classification. (Metallinou et al., 2008) and (Eyben et al., 2010a) fused audio and textual modalities for emotion recognition. Both approaches relied on a feature-level fusion. (Wu and Liang, 2011) fused audio and textual clues at decision level.

3 Method

In this work, we propose a LSTM network that takes as input the sequence of utterances in a video and extracts contextual unimodal and multimodal features by modeling the dependencies among the input utterances. M number of videos, comprising of its constituent utterances, serve as the input. We represent the dataset as $U = u_1, u_2, u_3..., u_M$ and each $u_i = u_{i,1}, u_{i,2}, ..., u_i, L_i$ where L_i is the number of utterances in video u_i. Below, we present an overview of the proposed method in two major steps.

A. Context-Independent Unimodal Utterance-Level Feature Extraction

Firstly, the unimodal features are extracted without considering the contextual information of the utterances (Section 3.1).

B. Contextual Unimodal and Multimodal Classification

Secondly, the context-independent unimodal features (from Step A) are fed into a LSTM network (termed contextual LSTM) that allows consecutive utterances in a video to share information in the feature extraction process (Section 3.2).

We experimentally show that this proposed framework improves the performance of utterance-level sentiment classification over traditional frameworks.

3.1 Extracting Context-Independent Unimodal Features

Initially, the unimodal features are extracted from each utterance separately, i.e., we do not consider the contextual relation and dependency among the utterances. Below, we explain the textual, audio, and visual feature extraction methods.

3.1.1 text-CNN: Textual Features Extraction

The source of textual modality is the transcription of the spoken words. For extracting features from the textual modality, we use a CNN (Karpathy et al., 2014). In particular, we first represent each utterance as the concatenation of vectors of the constituent words. These vectors are the publicly available 300-dimensional word2vec vectors trained on 100 billion words from Google News (Mikolov et al., 2013).

The convolution kernels are thus applied to these concatenated word vectors instead of individual words. Each utterance is wrapped to a window of 50 words which serves as the input to the CNN. The CNN has two convolutional layers; the first layer has two kernels of size 3 and 4, with 50 feature maps each and the second layer has a kernel of size 2 with 100 feature maps.

The convolution layers are interleaved with max-pooling layers of window 2×2. This is followed by a fully connected layer of size 500 and softmax output. We use a rectified linear unit (ReLU) (Teh and Hinton, 2001) as the activation function. The activation values of the fully-connected layer are taken as the features of utterances for text modality. The convolution of the CNN over the utterance learns abstract representations of the phrases equipped with implicit semantic information, which with each successive layer spans over increasing number of words and ultimately the entire utterance.

3.1.2 openSMILE: Audio Feature Extraction

Audio features are extracted at 30 Hz frame-rate and a sliding window of 100 ms. To compute the features, we use openSMILE (Eyben et al., 2010b), an open-source software that automatically extracts audio features such as pitch and voice intensity. Voice normalization is performed and voice intensity is thresholded to identify samples with and without voice. Z-standardization is used to perform voice normalization.

The features extracted by openSMILE consist of several low-level descriptors (LLD), e.g., MFCC, voice intensity, pitch, and their statistics, e.g., mean, root quadratic mean, etc. Specifically, we use IS13-ComParE configuration file in openSMILE. Taking into account all functionals of each LLD, we obtained 6373 features.

3.1.3 3D-CNN: Visual Feature Extraction

We use 3D-CNN (Ji et al., 2013) to obtain visual features from the video. We hypothesize that 3D-CNN will not only be able to learn relevant features from each frame, but will also learn the changes among given number of consecutive frames.

In the past, 3D-CNN has been successfully applied to object classification on tridimensional data (Ji et al., 2013). Its ability to achieve state-of-the-art results motivated us to adopt it in our framework.

Let $vid \in \mathbb{R}^{c \times f \times h \times w}$ be a video, where c = number of channels in an image (in our case $c = 3$, since we consider only RGB images), f = number of frames, h = height of the frames, and w = width of the frames. Again, we consider the 3D convolutional filter $filt \in \mathbb{R}^{f_m \times c \times f_d \times f_h \times f_w}$, where f_m = number of feature maps, c = number of channels, f_d = number of frames (in other words depth of the filter), f_h = height of the filter, and f_w = width of the filter. Similar to 2D-CNN, $filt$ slides across video vid and generates output $convout \in \mathbb{R}^{f_m \times c \times (f-f_d+1) \times (h-f_h+1) \times (w-f_w+1)}$. Next, we apply max pooling to $convout$ to select only relevant features. The pooling will be applied only to the last three dimensions of the array $convout$.

In our experiments, we obtained best results with 32 feature maps (f_m) with the filter-size of $5 \times 5 \times 5$ (or $f_d \times f_h \times f_w$). In other words, the dimension of the filter is $32 \times 3 \times 5 \times 5 \times 5$ (or $f_m \times c \times f_d \times f_h \times f_w$). Subsequently, we apply max pooling on the output of convolution operation, with window-size being $3 \times 3 \times 3$. This is followed by a dense layer of size 300 and softmax. The activation values of this dense layer are finally used as the video features for each utterance.

3.2 Context-Dependent Feature Extraction

In sequence classification, the classification of each member is dependent on the other members. Utterances in a video maintain a sequence. We hypothesize that, within a video, there is a high probability of inter-utterance dependency with respect to their sentimental clues.

In particular, we claim that, when classifying one utterance, other utterances can provide important contextual information. This calls for a model which takes into account such inter-dependencies and the effect these might have on the target utterance. To capture this flow of informational triggers across utterances, we use a LSTM-based recurrent neural network (RNN) scheme (Gers, 2001).

3.2.1 Long Short-Term Memory

LSTM (Hochreiter and Schmidhuber, 1997) is a kind of RNN, an extension of conventional feed-forward neural network. Specifically, LSTM cells are capable of modeling long-range dependencies, which other traditional RNNs fail to do given the vanishing gradient issue. Each LSTM cell consists of an input gate i, an output gate o, and a forget gate f, to control the flow of information.

Current research (Zhou et al., 2016) indicates the benefit of using such networks to incorporate contextual information in the classification process. In our case, the LSTM network serves the purpose of context-dependent feature extraction by modeling relations among utterances. We term our architecture 'contextual LSTM'. We propose several architectural variants of it later in the paper.

3.2.2 Contextual LSTM Architecture

Let unimodal features have dimension k, each utterance is thus represented by a feature vector $\boldsymbol{x}_{i,t} \in \mathbb{R}^k$, where t represents the t^{th} utterance of the video i. For a video, we collect the vectors for all the utterances in it, to get $\boldsymbol{X_i} = [\boldsymbol{x}_{i,1}, \boldsymbol{x}_{i,2}, ..., \boldsymbol{x}_{i,L_i}] \in \mathbb{R}^{L_i \times k}$, where L_i represents the number of utterances in the video. This matrix X_i serves as the input to the LSTM. Figure 1 demonstrates the functioning of this LSTM module.

In the procedure, *getLstmFeatures(X_i)* of Algorithm 1, each of these utterance $x_{i,t}$ is passed through a LSTM cell using the equations mentioned in line 32 to 37. The output of the LSTM cell $h_{i,t}$ is then fed into a dense layer and finally into a softmax layer (line 38 to 39). The activations of the dense layer $z_{i,t}$ are used as the context-dependent features of contextual LSTM.

3.2.3 Training

The training of the LSTM network is performed using categorical cross-entropy on each utterance's softmax output per video, i.e.,

$$loss = -\frac{1}{(\sum_{i=1}^{M} L_i)} \sum_{i=1}^{M} \sum_{j=1}^{L_i} \sum_{c=1}^{C} y_{i,c}^{j} \log_2(\hat{y}_{i,c}^{j}),$$

where M = total number of videos, L_i = number of utterances for i^{th} video, $y_{i,c}^{j}$ = original output of class c, and $\hat{y}_{i,c}^{j}$ = predicted output for j^{th} utterance of i^{th} video.

As a regularization method, dropout between the LSTM cell and dense layer is introduced to avoid overfitting. As the videos do not have the same number of utterances, padding is introduced to serve as neutral utterances. To avoid the proliferation of noise within the network, bit masking is done on these padded utterances to eliminate their effect in the network. Hyper-parameters tuning is done on the training set by splitting it into train and validation components with $80/20\%$ split.

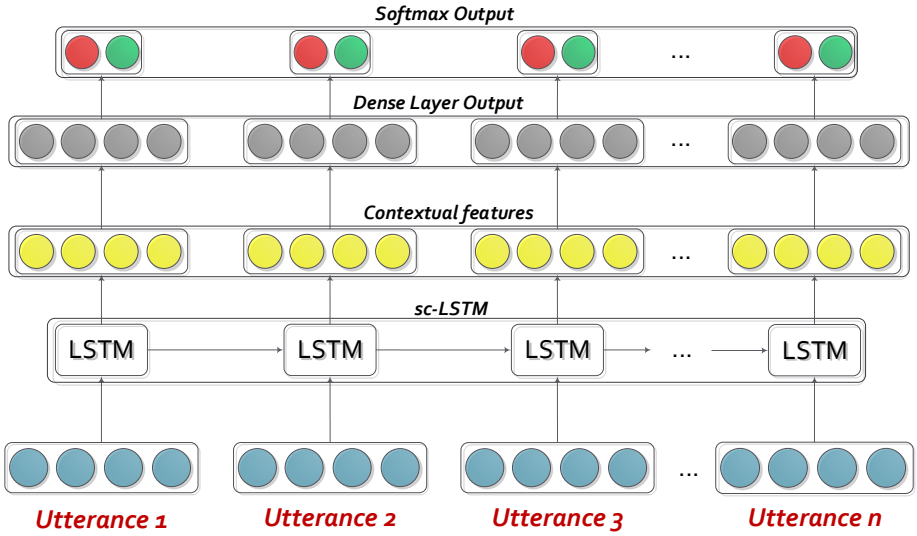

Figure 1: Contextual LSTM network: input features are passed through an unidirectional LSTM layer, followed by a dense and then a softmax layer. The dense layer activations serve as the output features.

RMSprop has been used as the optimizer which is known to resolve Adagrad's radically diminishing learning rates (Duchi et al., 2011). After feeding the training set to the network, the test set is passed through it to generate their context-dependent features. These features are finally passed through an SVM for the final classification.

Different Network Architectures We consider the following variants of the contextual LSTM architecture in our experiments.

sc-LSTM This variant of the contextual LSTM architecture consists of unidirectional LSTM cells. As this is the simple variant of the contextual LSTM, we termed it as simple contextual LSTM (sc-LSTM[1]).

h-LSTM We also investigate an architecture where the dense layer after the LSTM cell is omitted. Thus, the output of the LSTM cell $h_{i,t}$ provides our context-dependent features and the softmax layer provides the classification. We call this architecture hidden-LSTM (h-LSTM).

bc-LSTM Bi-directional LSTMs are two unidirectional LSTMs stacked together having opposite directions. Thus, an utterance can get information from utterances occurring before and after itself in the video. We replaced the regular LSTM with a bi-directional LSTM and named the resulting architecture as bi-directional contextual LSTM (bc-LSTM). The training process of this architecture is similar to sc-LSTM.

[1] http://github.com/senticnet/sc-lstm

uni-SVM In this setting, we first obtain the unimodal features as explained in Section 3.1, concatenate them and then send to an SVM for the final classification. It should be noted that using a gated recurrent unit (GRU) instead of LSTM did not improve the performance.

3.3 Fusion of Modalities

We accomplish multimodal fusion through two different frameworks, described below.

3.3.1 Non-hierarchical Framework

In this framework, we concatenate context-independent unimodal features (from Section 3.1) and feed that into the contextual LSTM networks, i.e., sc-LSTM, bc-LSTM, and h-LSTM.

3.3.2 Hierarchical Framework

Contextual unimodal features can further improve performance of the multimodal fusion framework explained in Section 3.3.1. To accomplish this, we propose a hierarchical deep network which consists of two levels.

Level-1 Context-independent unimodal features (from Section 3.1) are fed to the proposed LSTM network to get *context-sensitive* unimodal feature representations for each utterance. Individual LSTM networks are used for each modality.

Level-2 This level consists of a contextual LSTM network similar to Level-1 but independent in training and computation. Output from each LSTM network in Level-1 are concatenated and fed into this LSTM network, thus providing an inherent fusion scheme (see Figure 2).

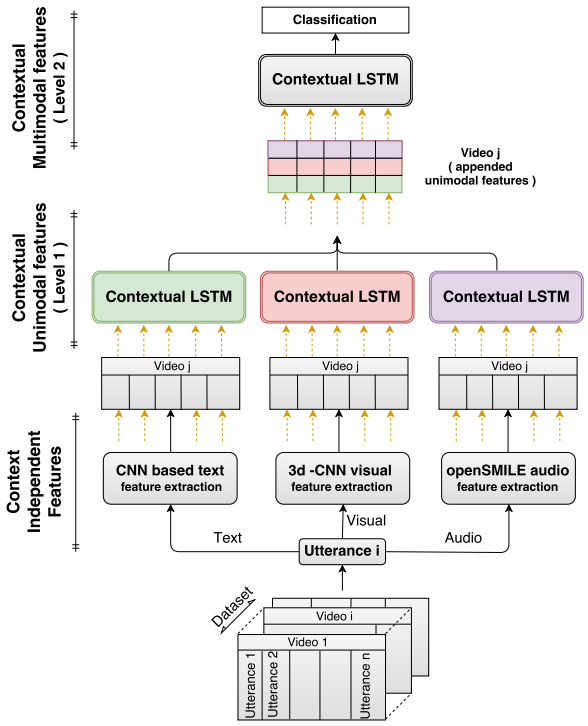

Figure 2: Hierarchical architecture for extracting context-dependent multimodal utterance features (see Figure 1 for the LSTM module).

The performance of the second level banks on the quality of the features from the previous level, with better features aiding the fusion process. Algorithm 1 describes the overall computation for utterance classification. For the hierarchical framework, we train Level-1 and Level-2 successively but separately, i.e., the training is not performed "end-to-end".

Weight		Bias	
W_i, W_f, W_c, W_o	$\in \mathbb{R}^{d \times k}$	b_i, b_f, b_c, b_o	$\in \mathbb{R}^{d}$
$P_i, P_f, P_c, P_o V_o$	$\in \mathbb{R}^{d \times d}$	b_z	$\in \mathbb{R}^{m}$
W_z	$\in \mathbb{R}^{m \times d}$	b_{sft}	$\in \mathbb{R}^{c}$
W_{sft}	$\in \mathbb{R}^{c \times m}$		

Table 1: Summary of notations used in Algorithm 1. Legenda: d = dimension of hidden unit; k = dimension of input vectors to LSTM layer; c = number of classes.

4 Experiments

4.1 Dataset details

Most of the research in multimodal sentiment analysis is performed on datasets with speaker overlap in train and test splits. Because each individual has a unique way of expressing emotions and sentiments, however, finding generic, person-independent features for sentiment analysis is very important.

Algorithm 1 Proposed Architecture

procedure TRAINARCHITECTURE(U, V)
  Train context-independent models with U
  for `i:[1,M]` **do** ▷ extract baseline features
    for `j:[1,`L_i`]` **do**
      $x_{i,j} \leftarrow TextFeatures(u_{i,j})$
      $x^{'}_{i,j} \leftarrow VideoFeatures(u_{i,j})$
      $x^{"}_{i,j} \leftarrow AudioFeatures(u_{i,j})$

  Unimodal:
  Train LSTM at Level-1 with $X, X^{'} and X^{"}$.
  for `i:[1,`M`]` **do** ▷ unimodal features
    $Z_i \leftarrow getLSTMFeatures(X_i)$
    $Z^{'}_i \leftarrow getLSTMFeatures(X^{'}_i)$
    $Z^{"}_i \leftarrow getLSTMFeatures(X^{"}_i)$

  Multimodal:
  for `i:[1,M]` **do**
    for `j:[1,`L_i`]` **do**
      if `Non-hierarchical fusion` **then**
        $x^{*}_{i,j} \leftarrow (x_{i,j} || x^{'}_{i,j} || x^{"}_{i,j})$ ▷ concatenation
      else
        if `Hierarchical fusion` **then**
          $x^{*}_{i,j} \leftarrow (z_{i,j} || z^{'}_{i,j} || z^{"}_{i,j})$ ▷ concatenation
  Train LSTM at Level-2 with X^*.
  for `i:[1,`M`]` **do** ▷ multimodal features
    $Z^{*}_i \leftarrow getLSTMFeatures(X^{*}_i)$
  testArchitecture(V)
  return Z^*
procedure TESTARCHITECTURE(V)
  Similar to training phase. V is passed through the learnt models to get the features and classification outputs. Table 1 shows the trainable parameters.
procedure GETLSTMFEATURES(X_i) ▷ for i^{th} video
  $Z_i \leftarrow \phi$
  for `t:[1,`L_i`]` **do** ▷ Table 1 provides notation
    $i_t \leftarrow \sigma(W_i x_{i,t} + P_i . h_{t-1} + b_i)$
    $\widetilde{C_t} \leftarrow tanh(W_c x_{i,t} + P_c h_{t-1} + b_c)$
    $f_t \leftarrow \sigma(W_f x_t + P_f h_{t-1} + b_f)$
    $C_t \leftarrow i_t * \widetilde{C_t} + f_t * C_{t-1}$
    $o_t \leftarrow \sigma(W_o x_t + P_o h_{t-1} + V_o C_t + b_o)$
    $h_t \leftarrow o_t * tanh(C_t)$ ▷ output of lstm cell
    $z_t \leftarrow ReLU(W_z h_t + b_z)$ ▷ dense layer
    $prediction \leftarrow softmax(W_{sft} z_t + b_{sft})$
    $Z_i \leftarrow Z_i \cup z_t$
  return Z_i

In real-world applications, the model should be robust to person idiosyncrasy but it is very difficult to come up with a generalized model from the behavior of a limited number of individuals. To this end, we perform person-independent experiments to study generalization of our model, i.e., our train/test splits of the datasets are completely disjoint with respect to speakers.

Multimodal Sentiment Analysis Datasets

MOSI The MOSI dataset (Zadeh et al., 2016) is a dataset rich in sentimental expressions where 93 people review topics in English. The videos

are segmented with each segments sentiment label scored between +3 (strong positive) to -3 (strong negative) by 5 annotators. We took the average of these five annotations as the sentiment polarity and, hence, considered only two classes (positive and negative). The train/validation set consists of the first 62 individuals in the dataset. The test set contains opinionated videos by rest 31 speakers. In particular, 1447 and 752 utterances are used in training and test, respectively.

MOUD This dataset (Pérez-Rosas et al., 2013) contains product review videos provided by 55 persons. The reviews are in Spanish (we used Google Translate API[2] to get the English transcripts). The utterances are labeled to be either *positive, negative or neutral*. However, we drop the *neutral* label to maintain consistency with previous work. Out of 79 videos in the dataset, 59 videos are considered in the train/val set.

Multimodal Emotion Recognition Datasets

IEMOCAP The IEMOCAP (Busso et al., 2008) contains the acts of 10 speakers in a two-way conversation segmented into utterances. The medium of the conversations in all the videos is English. The database contains the following categorical labels: anger, happiness, sadness, neutral, excitement, frustration, fear, surprise, and other, but we take only the first four so as to compare with the state of the art (Rozgic et al., 2012). Videos by the first 8 speakers are considered in the training set. The train/test split details are provided in Table 2, which provides information regarding train/test split of all the datasets. Table 2 also provides cross-dataset split details where the datasets MOSI and MOUD are used for training and testing, respectively. The proposed model being used on reviews from different languages allows us to analyze its robustness and generalizability.

4.1.1 Characteristic of the Datasets

In order to evaluate the robustness of our proposed method, we employ it on multiple datasets of different kinds. Both MOSI and MOUD are used for the sentiment classification task but they consist of review videos spoken in different languages, i.e., English and Spanish, respectively.

[2] http://translate.google.com

IEMOCAP dataset is different from MOSI and MOUD since it is annotated with emotion labels. Apart from this, IEMOCAP dataset was created using a different method than MOSI and MOUD. These two datasets were developed by crawling consumers' spontaneous online product review videos from popular social websites and later labeled with sentiment labels. To curate the IEMOCAP dataset, instead, subjects were provided affect-related scripts and asked to act.

As pointed out by Poria et al. (Poria et al., 2017a), acted dataset like IEMOCAP can suffer from biased labeling and incorrect acting which can further cause the poor generalizability of the models trained on the acted datasets.

Dataset	Train		Test	
	uttrnce	*video*	*uttrnce*	*video*
IEMOCAP	4290	120	1208	31
MOSI	1447	62	752	31
MOUD	322	59	115	20
MOSI → MOUD	2199	93	437	79

Table 2: uttrnce: Utterance; Person-Independent Train/Test split details of each dataset (≈ 70/30 % split). Legenda: X→Y represents train: X and test: Y; Validation sets are extracted from the shuffled training sets using 80/20 % train/val ratio.

It should be noted that the datasets' individual configuration and splits are same throughout all the experiments (i.e., context-independent unimodal feature extraction, LSTM-based context-dependent unimodal and multimodal feature extraction and classification).

4.2 Performance of Different Models

In this section, we present unimodal and multimodal sentiment analysis performance of different LSTM network variants as explained in Section 3.2.3 and comparison with the state of the art.

Hierarchical vs Non-hierarchical Fusion Framework As expected, trained contextual unimodal features help the hierarchical fusion framework to outperform the non-hierarchical framework. Table 3 demonstrates this by comparing the hierarchical and the non-hierarchical frameworks using the bc-LSTM network.

For this reason, we the rest of the analysis only leverages on the hierarchical framework. The non-hierarchical model outperforms the baseline uni-SVM, which confirms that it is the context-sensitive learning paradigm that plays the key role in improving performance over the baseline.

Comparison of Different Network Variants It is to be noted that both sc-LSTM and bc-LSTM perform quite well on the multimodal emotion recognition and sentiment analysis datasets. Since bc-LSTM has access to both the preceding and following information of the utterance sequence, it performs consistently better on all the datasets over sc-LSTM. The usefulness of the dense layer in increasing the performance is evident from the experimental results shown in Table 3. The performance improvement is in the range of 0.3% to 1.5% on MOSI and MOUD datasets. On the IEMOCAP dataset, the performance improvement of bc-LSTM and sc-LSTM over h-LSTM is in the range of 1% to 5%.

Comparison with the Baselines Every LSTM network variant has outperformed the baseline uni-SVM on all the datasets by the margin of 2% to 5% (see Table 3). These results prove our initial hypothesis that modeling the contextual dependencies among utterances (which uni-SVM cannot do) improves the classification. The higher performance improvement on the IEMOCAP dataset indicates the necessity of modeling long-range dependencies among the utterances as continuous emotion recognition is a multiclass sequential problem where a person does not frequently change emotions (Wöllmer et al., 2008). We have implemented and compared with the current state-of-the-art approach proposed by (Poria et al., 2015). In their method, they extracted features from each modality and fed these to a MKL classifier. However, they did not conduct the experiment in a speaker-independent manner and also did not consider the contextual relation among the utterances. In Table 3, the results in bold are statistically significant ($p < 0.05$) in compare to uni-SVM. Experimental results in Table 4 show that the proposed method outperformes (Poria et al., 2015) by a significant margin. For the emotion recognition task, we have compared our method with the current state of the art (Rozgic et al., 2012), who extracted features in a similar fashion to (Poria et al., 2015) (although they used SVM trees (Yuan et al., 2006) for the fusion).

4.3 Importance of the Modalities

As expected, in all kinds of experiments, bimodal and trimodal models have outperformed unimodal models. Overall, audio modality has performed better than visual on all the datasets.

On MOSI and IEMOCAP datasets, the textual classifier achieves the best performance over other unimodal classifiers. On IEMOCAP dataset, the unimodal and multimodal classifiers obtained poor performance to classify *neutral* utterances. The textual modality, combined with non-textual modes, boosts the performance in IEMOCAP by a large margin. However, the margin is less in the other datasets.

On the MOUD dataset, the textual modality performs worse than audio modality due to the noise introduced in translating Spanish utterances to English. Using Spanish word vectors[3] in *text-CNN* results in an improvement of 10%. Nonetheless, we report results using these translated utterances as opposed to utterances trained on Spanish word vectors, in order to make fair comparison with (Poria et al., 2015).

4.4 Generalization of the Models

To test the generalizability of the models, we have trained our framework on complete MOSI dataset and tested on MOUD dataset (Table 5). The performance was poor for audio and textual modality as the MOUD dataset is in Spanish while the model is trained on MOSI dataset, which is in English language. However, notably the visual modality performs better than the other two modalities in this experiment, which means that in cross-lingual scenarios facial expressions carry more generalized, robust information than audio and textual modalities. We could not carry out a similar experiment for emotion recognition as no other utterance-level dataset apart from the IEMOCAP was available at the time of our experiments.

4.5 Qualitative Analysis

The need for considering context dependency (see Section 1) is of prime importance for utterance-level sentiment classification. For example, in the utterance *"What would have been a better name for the movie"*, the speaker is attempting to comment the quality of the movie by giving an appropriate name. However, the sentiment is expressed implicitly and requires the contextual knowledge about the mood of the speaker and his/her general opinion about the film. The baseline unimodal-SVM and state of the art fail to classify this utterance correctly[4].

[3] http://crscardellino.me/SBWCE

[4] RNTN classifies it as neutral. It can be seen here http://nlp.stanford.edu:8080/sentiment/rntnDemo.html

Modality	MOSI					MOUD					IEMOCAP				
	hierarchical (%)				non-hier (%)	hierarchical (%)				non-hier (%)	hierarchical (%)				non-hier (%)
	uni-SVM	h-LSTM	sc-LSTM	bc-LSTM		uni-SVM	h-LSTM	sc-LSTM	bc-LSTM		uni-SVM	h-LSTM	sc-LSTM	bc-LSTM	
T	75.5	77.4	77.6	**78.1**		49.5	50.1	51.3	**52.1**		65.5	68.9	71.4	**73.6**	
V	53.1	55.2	55.6	**55.8**		46.3	48.0	48.2	**48.5**		47.0	52.0	52.6	**53.2**	
A	58.5	59.6	59.9	**60.3**		51.5	56.3	57.5	**59.9**		52.9	54.4	55.2	**57.1**	
T + V	76.7	78.9	79.9	**80.2**	78.5	50.2	50.6	51.3	**52.2**	50.9	68.5	70.3	72.3	**75.4**	73.2
T + A	75.8	78.3	78.8	**79.3**	78.2	53.1	56.9	57.4	**60.4**	55.5	70.1	74.1	75.2	**75.6**	74.5
V + A	58.6	61.5	61.8	**62.1**	60.3	62.8	62.9	64.4	**65.3**	64.2	67.6	67.8	68.2	**68.9**	67.3
T + V + A	77.9	78.1	78.6	**80.3**	78.1	66.1	66.4	67.3	**68.1**	67.0	72.5	73.3	74.2	**76.1**	73.5

Table 3: Comparison of models mentioned in Section 3.2.3. The table reports the accuracy of classification. Legenda: non-hier ← Non-hierarchical bc-lstm. For remaining fusion, hierarchical fusion framework is used (Section 3.3.2).

Modality	Sentiment (%)		Emotion on IEMOCAP (%)			
	MOSI	MOUD	*angry*	*happy*	*sad*	*neutral*
T	78.12	52.17	76.07	78.97	76.23	67.44
V	55.80	48.58	53.15	58.15	55.49	51.26
A	60.31	59.99	58.37	60.45	61.35	52.31
T + V	80.22	52.23	77.24	78.99	78.35	68.15
T + A	79.33	60.39	77.15	79.10	78.10	69.14
V + A	62.17	65.36	68.21	71.97	70.35	62.37
A + V + T	**80.30**	**68.11**	**77.98**	**79.31**	**78.30**	**69.92**
State-of -the-art	73.55[1]	63.25[1]	73.10 [2]	72.40[2]	61.90[2]	58.10[2]

[1]by (Poria et al., 2015),[2]by (Rozgic et al., 2012)

Table 4: Accuracy % on textual (T), visual (V), audio (A) modality and comparison with the state of the art. For the fusion, the hierarchical fusion framework was used.

Modality	MOSI → MOUD			
	uni-SVM	h-LSTM	sc-LSTM	bc-LSTM
T	46.5%	46.5%	46.6%	**46.9%**
V	43.3%	45.5%	48.3%	**49.6%**
A	42.9%	46.0%	46.4%	**47.2%**
T + V	49.8%	49.8%	49.8%	**49.8%**
T + A	50.4%	50.9%	51.1%	**51.3%**
V + A	46.0%	47.1%	49.3%	**49.6%**
T + V + A	51.1%	52.2%	52.5%	**52.7%**

Table 5: Cross-dataset comparison in terms of classification accuracy.

However, information from neighboring utterances, e.g., *"And I really enjoyed it"* and *"The countryside which they showed while going through Ireland was astoundingly beautiful"* indicate its positive context and help our contextual model to classify the target utterance correctly. Such contextual relationships are prevalent throughout the dataset.

In order to have a better understanding of the roles of each modality for the overall classification, we have also done some qualitative analysis. For example, the utterance *"who doesn't have any presence or greatness at all"* was classified as positive by the audio classifier (as "presence and greatness at all" was spoken with enthusiasm). However, the textual modality caught the negation induced by "doesn't" and classified it correctly. The same happened to the utterance "amazing special effects", which presented no jest of enthusiasm in the speaker's voice nor face, but was correctly classified by the textual classifier.

On other hand, the textual classifier classified the utterance "that like to see comic book characters treated responsibly" as positive (for the presence of "like to see" and "responsibly") but the high pitch of anger in the person's voice and the frowning face helps to identify this as a negative utterance. In some cases, the predictions of the proposed method are wrong because of face occlusion or noisy audio. Also, in cases where sentiment is very weak and non contextual, the proposed approach shows some bias towards its surrounding utterances, which further leads to wrong predictions.

5 Conclusion

The contextual relationship among utterances in a video is mostly ignored in the literature. In this paper, we developed a LSTM-based network to extract contextual features from the utterances of a video for multimodal sentiment analysis. The proposed method has outperformed the state of the art and showed significant performance improvement over the baseline.

As future work, we plan to develop a LSTM-based attention model to determine the importance of each utterance and its specific contribution to each modality for sentiment classification.

References

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

Erik Cambria, Dipankar Das, Sivaji Bandyopadhyay, and Antonio Feraco. 2017a. *A Practical Guide to Sentiment Analysis*. Springer, Cham, Switzerland.

Erik Cambria, Devamanyu Hazarika, Soujanya Poria, Amir Hussain, and RBV Subramanyam. 2017b. Benchmarking multimodal sentiment aanlysis. In *CICLing*.

Erik Cambria, Soujanya Poria, Rajiv Bajpai, and Björn Schuller. 2016. SenticNet 4: A semantic resource for sentiment analysis based on conceptual primitives. In *COLING*. pages 2666–2677.

Erik Cambria and Bebo White. 2014. Jumping NLP curves: A review of natural language processing research. *IEEE Computational Intelligence Magazine* 9(2):48–57.

Lawrence S Chen, Thomas S Huang, Tsutomu Miyasato, and Ryohei Nakatsu. 1998. Multimodal human emotion/expression recognition. In *Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, pages 366–371.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research* 12(Aug):2493–2537.

Dragos Datcu and L Rothkrantz. 2008. Semantic audio-visual data fusion for automatic emotion recognition. *Euromedia'2008* .

Liyanage C De Silva, Tsutomu Miyasato, and Ryohei Nakatsu. 1997. Facial emotion recognition using multi-modal information. In *Proceedings of ICICS*. IEEE, volume 1, pages 397–401.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research* 12(Jul):2121–2159.

Paul Ekman. 1974. Universal facial expressions of emotion. *Culture and Personality: Contemporary Readings/Chicago* .

Florian Eyben, Martin Wöllmer, Alex Graves, Björn Schuller, Ellen Douglas-Cowie, and Roddy Cowie. 2010a. On-line emotion recognition in a 3-d activation-valence-time continuum using acoustic and linguistic cues. *Journal on Multimodal User Interfaces* 3(1-2):7–19.

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

Felix Gers. 2001. *Long Short-Term Memory in Recurrent Neural Networks*. Ph.D. thesis, Universität Hannover.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 2013. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence* 35(1):221–231.

Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. pages 1725–1732.

Loic Kessous, Ginevra Castellano, and George Caridakis. 2010. Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis. *Journal on Multimodal User Interfaces* 3(1-2):33–48.

Yukun Ma, Erik Cambria, and Sa Gao. 2016. Label embedding for zero-shot fine-grained named entity typing. In *COLING*. Osaka, pages 171–180.

Navonil Majumder, Soujanya Poria, Alexander Gelbukh, and Erik Cambria. 2017. Deep learning based document modeling for personality detection from text. *IEEE Intelligent Systems* 32(2):74–79.

Angeliki Metallinou, Sungbok Lee, and Shrikanth Narayanan. 2008. Audio-visual emotion recognition using gaussian mixture models for face and voice. In *Tenth IEEE International Symposium on ISM 2008*. IEEE, pages 250–257.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* .

David Olson. 1977. From utterance to text: The bias of language in speech and writing. *Harvard educational review* 47(3):257–281.

Luca Oneto, Federica Bisio, Erik Cambria, and Davide Anguita. 2016. Statistical learning theory and ELM for big social data analysis. *IEEE Computational Intelligence Magazine* 11(3):45–55.

Verónica Pérez-Rosas, Rada Mihalcea, and Louis-Philippe Morency. 2013. Utterance-level multimodal sentiment analysis. In *ACL (1)*. pages 973–982.

Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017a. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion* .

Soujanya Poria, Erik Cambria, and Alexander Gelbukh. 2015. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *Proceedings of EMNLP*. pages 2539–2544.

Soujanya Poria, Erik Cambria, and Alexander Gelbukh. 2016a. Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems* 108:42–49.

Soujanya Poria, Erik Cambria, D Hazarika, and Prateek Vij. 2016b. A deeper look into sarcastic tweets using deep convolutional neural networks. In *COLING*. pages 1601–1612.

Soujanya Poria, Iti Chaturvedi, Erik Cambria, and Federica Bisio. 2016c. Sentic LDA: Improving on LDA with semantic similarity for aspect-based sentiment analysis. In *IJCNN*. pages 4465–4473.

Soujanya Poria, Iti Chaturvedi, Erik Cambria, and Amir Hussain. 2016d. Convolutional mkl based multimodal emotion recognition and sentiment analysis. In *Data Mining (ICDM), 2016 IEEE 16th International Conference on*. IEEE, pages 439–448.

Soujanya Poria, Haiyun Peng, Amir Hussain, Newton Howard, and Erik Cambria. 2017b. Ensemble application of convolutional neural networks and multiple kernel learning for multimodal sentiment analysis. *Neurocomputing* .

Dheeraj Rajagopal, Erik Cambria, Daniel Olsher, and Kenneth Kwok. 2013. A graph-based approach to commonsense concept extraction and semantic similarity detection. In *WWW*. Rio De Janeiro, pages 565–570.

Viktor Rozgic, Sankaranarayanan Ananthakrishnan, Shirin Saleem, Rohit Kumar, and Rohit Prasad. 2012. Ensemble of svm trees for multimodal emotion recognition. In *Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific*. IEEE, pages 1–4.

Björn Schuller. 2011. Recognizing affect from linguistic information in 3d continuous space. *IEEE Transactions on Affective Computing* 2(4):192–205.

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of EMNLP*. pages 1631–1642.

Vee Teh and Geoffrey E Hinton. 2001. Rate-coded restricted boltzmann machines for face recognition. In T Leen, T Dietterich, and V Tresp, editors, *Advances in neural information processing system*. volume 13, pages 908–914.

Martin Wöllmer, Florian Eyben, Stephan Reiter, Björn W Schuller, Cate Cox, Ellen Douglas-Cowie, Roddy Cowie, et al. 2008. Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies. In *Interspeech*. volume 2008, pages 597–600.

Martin Wollmer, Felix Weninger, Timo Knaup, Bjorn 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.

Chung-Hsien Wu and Wei-Bin Liang. 2011. Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. *IEEE Transactions on Affective Computing* 2(1):10–21.

Xun Yuan, Wei Lai, Tao Mei, Xian-Sheng Hua, Xiu-Qing Wu, and Shipeng Li. 2006. Automatic video genre categorization using hierarchical svm. In *Image Processing, 2006 IEEE International Conference on*. IEEE, pages 2905–2908.

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

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *The 54th Annual Meeting of the Association for Computational Linguistics*. pages 207–213.