Edison Marrese-Taylor, Cristian Rodriguez-Opazo, Jorge A. Balazs, Stephen Gould and Yutaka Matsuo
A Multi-modal Approach to Fine-grained Opinion Mining on Video Reviews
Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020), 2020, pp.8-18

Published version http://dx.doi.org/10.18653/v1/2020.challengehml-1.2

A Multi-modal Approach to Fine-grained Opinion Mining on Video Reviews

Edison Marrese-Taylor[1*], **Cristian Rodriguez-Opazo**[2*], **Jorge A. Balazs**[1]
Stephen Gould[2] and **Yutaka Matsuo**[1]
Graduate School of Engineering, The University of Tokyo, Japan[1]
{emarrese, jorge, matsuo}@weblab.t.u-tokyo.ac.jp
Australian Centre for Robotic Vision (ACRV), Australian National University[2]
{cristian.rodriguez, stephen.gould}@anu.edu.au
[*]Authors contributed equally to this work.

Abstract

Despite the recent advances in opinion mining for written reviews, few works have tackled the problem on other sources of reviews. In light of this issue, we propose a multi-modal approach for mining fine-grained opinions from video reviews that is able to determine the aspects of the item under review that are being discussed and the sentiment orientation towards them. Our approach works at the sentence level without the need for time annotations and uses features derived from the audio, video and language transcriptions of its contents. We evaluate our approach on two datasets and show that leveraging the video and audio modalities consistently provides increased performance over text-only baselines, providing evidence these extra modalities are key in better understanding video reviews.

1 Introduction

Sentiment analysis (SA) is an important task in natural language processing, aiming at identifying and extracting opinions, emotions, and subjectivity. As a result, sentiment can be automatically collected, analyzed and summarized. Because of this, SA has received much attention not only in academia but also in industry, helping provide feedback based on customers' opinions about products or services. The underlying assumption in SA is that the entire input has an overall polarity, however, this is usually not the case. For example, laptop reviews generally not only express the overall sentiment about a specific model (e.g., "This is a great laptop"), but also relate to its specific aspects, such as the hardware, software or price. Subsequently, a review may convey opposing sentiments (e.g., "Its performance is ideal, I wish I could say the same about the price") or objective information (e.g., "This one still has the CD slot") for different aspects of an entity. Aspect-based sentiment analysis (ABSA) or fine-grained opinion mining aims to extract opinion targets or aspects of entities being reviewed in a text, and to determine the sentiment reviewers express for each. ABSA allows us to evaluate aggregated sentiments for each aspect of a given product or service and gain a more granular understanding of their quality. This is of especial interest for companies as it enables them to refine specifications for a given product or service, and leading to an improved overall customer satisfaction.

Fine-grained opinion mining is also important for a variety of NLP tasks, including opinion-oriented question answering and opinion summarization. In practical terms, the ABSA task can be divided into two sub-steps, namely aspect extraction (*AE*) and (aspect level) sentiment classification (*SC*), which can be tackled in a pipeline fashion, or simultaneously (*AESC*). These tasks can be regarded as a token-level sequence labeling problem, and are generally tackled using supervised learning. The 2014 and 2015 SemEval workshops, co-located with COLING 2014 and NAACL 2015 respectively, included shared tasks on ABSA (Pontiki et al., 2014) and also followed this approach, which has also served as a way to encourage developments alongside this line of research (Mitchell et al., 2013; Irsoy and Cardie, 2014; Liu et al., 2015; Zhang et al., 2015).

The flexibility provided by the deep learning setting has helped multi-modal approaches to bloom. Examples of this include tasks such as machine translation (Specia et al., 2016; Elliott et al., 2017), word sense disambiguation (Chen et al., 2015), visual question answering (Chen et al., 2017), language grounding (Beinborn et al.; Lazaridou et al., 2015), and sentiment analysis (Poria et al., 2015; Zadeh et al., 2016). Specifically in this last example, the task focuses on generalizing text-based sentiment analysis to opinionated videos, where three communicative modalities are present: language (spoken words), visual (gestures), and acoustic (voice).

Although reviews often come under the form of a written commentary, people are increasingly turning to video platforms such as YouTube looking for product reviews to help them shop. In this context, Marrese-Taylor et al. (2017) explored a new direction, arguing that video reviews are the natural evolution of written product reviews and introduced a dataset of annotated video product review transcripts. Similarly, Garcia et al. (2019b) recently presented an improved version of the POM movie review dataset (Park et al., 2014),

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8–18
Seattle, USA, July5 - 10, 2020.
https://doi.org/10.18653/v1/P17

with annotated fine-grained opinions.

Although the videos in these kinds of datasets represent a rich multi-modal source of opinions, the features of the language in them may fundamentally differ from written reviews given that information is conveyed through multiple channels (one for speech, one for gestures, one for facial expressions, one for vocal inflections, etc.) In these, different information channels complement each other to maximize the coherence and clarity of their message. This means that although the content of each channel may be comprehended in isolation, in theory we need to process the information in all the channels simultaneously to fully comprehend the message (Hasan et al., 2019). In this context, information extracted from nonverbal language in videos, such as gestures and facial expressions, as well as from audio in the manner of voice inflections or pauses, and from scenes, object or images in the video, become critical for performing well.

In light of this, our paper introduces a multi-modal approach for fine-grained opinion mining. We conduct extensive experiments on two datasets built upon transcriptions of video reviews, Youtubean (Marrese-Taylor et al., 2017) and a fine-grain annotated version of the Persuasive Opinion Multimedia (POM) dataset (Park et al., 2014; Garcia et al., 2019b), adapting them to our setting by associating timestamps to each annotated sentence using the video subtitles. Our results demonstrate the effectiveness of our proposed approach and show that by leveraging the additional modalities we can consistently obtain better performance.

2 Related Work

Our work is related to aspect extraction using deep learning, a task that is often tackled as a sequence labeling problem. In particular, our work is related to Irsoy and Cardie (2014), who pioneered in the field by using multi-layered RNNs. Later, Liu et al. (2015) successfully adapted the architectures by Mesnil et al. (2013) which were originally developed for slot-filling in the context of Natural Language Understanding.

Literature offers related work on the usage of RNNs for open domain targeted sentiment (Mitchell et al., 2013), where Zhang et al. (2015) experimented with neural CRF models using various RNN architectures on a dataset of informal language from Twitter.

Regarding target-based sentiment analysis, the literature contains several ad-hoc models that account for the sentence structure and the position of the aspect on it (Tang et al., 2016a,b). These approaches mainly use attention-augmented RNNs for solving the task. However, they require the location of the aspect to be known in advance and therefore are only useful in pipeline models, while instead we model aspect extraction and sentiment classification as a joint task or using multi-tasking.

AESC has also often been tackled as a sequence labeling problem, mainly using Conditional Random Fields (CRFs) (Mitchell et al., 2013). To model the problem in this fashion, collapsed or sentiment-bearing IOB labels (Zhang et al., 2015) are used. Pipeline models (i.e. task-independent model ensembles) have also been extensively studied by the same authors. Xu et al. (2014) performed *AESC* by modeling the linking relation between aspects and the sentiment-bearing phrases.

When it comes to the video review domain, there is related work on YouTube mining, mainly focused on exploiting user comments. For example, Wu et al. (2014) exploited crowdsourced textual data from time-synced commented videos, proposing a temporal topic model based on LDA. Tahara et al. (2010) introduced a similar approach for *Nico Nico*, using time-indexed social annotations to search for desirable scenes inside videos.

On the other hand, Severyn et al. (2014) proposed a systematic approach to mine user comments that relies on tree kernel models. Additionally, Krishna et al. (2013) performed sentiment analysis on YouTube comments related to popular topics using machine learning techniques, showing that the trends in users' sentiments is well correlated to the corresponding real-world events. Siersdorfer et al. (2010) presented an analysis of dependencies between comments and comment ratings, proving that community feedback in combination with term features in comments can be used for automatically determining the community acceptance of comments.

We also find some papers that have successfully attempted to use closed caption mining for video activity recognition (Gupta and Mooney, 2010) and scene segmentation (Gupta and Mooney, 2009). Similar work has been done using closed captions to classify movies by genre (Brezeale and Cook, 2006) and summarize video programs (Brezeale and Cook, 2006). Regarding multi-modal approaches for sentiment analysis, we see that previous work has focused mainly on sentiment classification, or the related task of emotion detection (Lakomkin et al., 2017), where the CMU MOSI dataset (Zadeh et al., 2016) appears as the main resource. In this setting, the main problem is how to model and capture cross-modality interactions to predict the sentiment correctly. In this regard Zadeh et al. (2017) proposed a tensor fusion layer that can better capture cross-modality interactions between text, audio and video inputs, while Poria et al. (2017) modeled inter-dependencies across difference utterances of a single video, obtaining further improvements.

Blanchard et al. (2018) are, to the best of our knowledge, the first to tackle scalable multi-modal sentiment classification using both visual and acoustic modalities. More recently Ghosal et al. (2018) proposed an RNN-based multi-modal approach that relies on attention to learn the contributing features among multi-utterance representations. On the other hand Pham et al. (2018) introduced multi-modal sequence-to-sequence models

which perform specially well in bi-modal settings. Finally, Akhtar et al. (2019) proposed a multi-modal, multi-task approach in which the inputs from a video (text, acoustic and visual frames), are exploited for simultaneously predicting the sentiment and expressed emotions of an utterance. Our work is related to all of these approaches, but it is different in that we apply multi-modal techniques not only for sentiment classification, but also for aspect extraction.

Finally, Marrese-Taylor et al. (2017) and Garcia et al. (2019b) contributed multi-modal datasets obtained from product and movie reviews respectively, specifically for the task of fine-grained opinion mining. Furthermore, Garcia et al. (2019a) recently used the latter to propose a hierarchical multi-modal model for opinion mining. Compared to them, our approach follows a more traditional setting for fine-grained opinion mining, while also offering a more general framework for the problem. Garcia et al. (2019a) utilize a single encoder that receives as input the concatenation of the features for each modality, for each token. This requires explicit alignment between the features of the different modalities at the token level. In contrast, since each modality is encoded separately in our approach, we only require the feature alignment to be at the sentence level.

3 Task Description

Opinion mining can be performed at several levels of granularity, the most common ones being the sentence level, and the more fine-grained aspect level. Fine-grained opinion mining can be further subdivided in two tasks: aspect extraction and aspect-level sentiment classification. The former deals with finding the aspects being referred to, and the latter with associating them with a sentiment.

Previous work usually casts this task as a sequence-labeling problem, where models have to predict whether a token is a part of an aspect and infer its sentiment polarity (Mitchell et al., 2013; Zhang et al., 2015; Liu et al., 2015). Depending on the dataset annotations, aspect categories are in some cases specified as well.

Formally, given a sentence $s = [x_1, \ldots, x_n]$, we want to automatically annotate each token x_i with its aspect membership and polarity. In the simpler case where we only want to perform Aspect Extraction, a common annotation scheme is to tag each token with a label $y_i \in \mathbb{L}^{\text{AE}}$ where $\mathbb{L}^{\text{AE}} = \{I, O, B\}$. In this scheme, commonly known as IOB, O labels indicate that a token is not a member of an aspect, B labels indicate that a token is at the beginning of an aspect, and I labels indicate that the token is inside an aspect.

Similarly, performing token-level Sentiment Classification only is equivalent to tagging each token with a label $y_i \in \mathbb{L}^{\text{SC}}$ where $\mathbb{L}^{\text{SC}} = \{\phi, +, -\}$, and ϕ denotes no sentiment, $+$ denotes a positive polarity and $-$ a negative one.

It is also possible to define a *collapsed* annotation scheme, where aspect membership and sentiment polarity are encoded in a single tag. We define the label set for this setting as $\mathbb{L}^{\text{C}} = \{O, B+, B-, I+, I-\}$.

Table 1 shows the possible ways to annotate the sentence "I love the saturated colors!" under these three annotation schemes, where the aspect being referred to is "saturated colors".

	I	**love**	**the**	**saturated**	**colors**	**!**
$\mathbb{L}^{\text{AE}}$	O	O	O	B	I	O
$\mathbb{L}^{\text{SC}}$	ϕ	ϕ	ϕ	$+$	$+$	ϕ
$\mathbb{L}^{\text{C}}$	O	O	O	$B+$	$I+$	O

Table 1: Label definition alternatives for the tasks in ABSA using sequence labeling.

Labels can be further augmented with type information. For example Liu et al. (2015) used different tags for opinion targets (e.g. B-TARG), and opinion expressions (e.g., B-EXPR), however, we do not rely on this information.

4 Proposed Approach

We propose a multi-modal approach for aspect extraction and sentiment classification that leverages video, audio and textual features. This approach assumes we have a video review v containing opinions, its extracted audio stream a, and a transcription of the audio into a sequence of sentences $\mathbb{S}$. Further, each sentence $s \in \mathbb{S}$ is annotated with its respective start and end times in the video effectively mapping them to a video segment $v^s \subset v$ and its corresponding audio segment $a^s \subset a$. These segments do not necessarily cover the whole video i.e. $\cup v^s \subset v$ since the reviews may include parts that have no speech and therefore no sentences are associated to those. Our end goal is to produce a sequence of labels $l = [y_1, \ldots, y_n]$ for each sentence $s = [x_1, \ldots, x_n]$ while exploiting the information contained in v^s and a^s.

Figure 1 presents a high-level overview of our approach. We rely on an encoder-decoder paradigm to create separate representations for each modality (Cho et al., 2014). The text encoding module generates a representation for each token in the input text, while the video and audio encoding layers produce utterance-level representations from each modality.

We propose combining these representations with an approach inspired by early-fusion (Xu et al., 2018), which allows for the word-level representations to interact with audio and visual features. Finally, a sequence labeling module is in charge of taking the final token-level representations and producing a token-level label. In the following sub-sections we describe each component of our model.

4.1 Text Encoding Module

This module generates a representation of the natural language input so that the obtained representation is

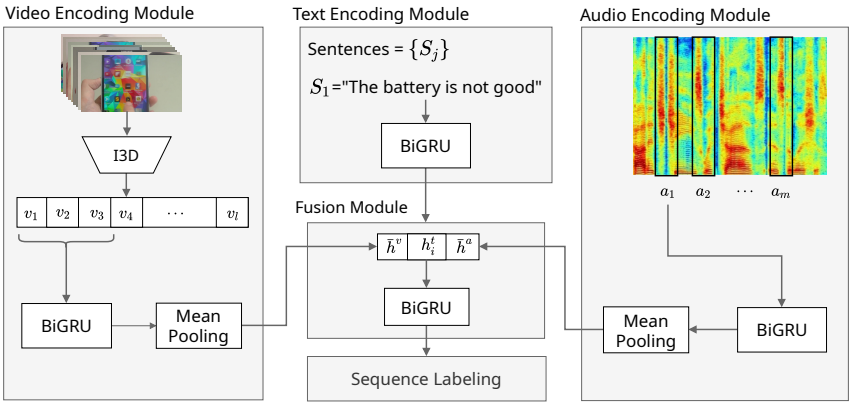

Figure 1: Overview of our proposed approach for multi-modal opinion mining

useful for the sequence labeling task. Our text encoder first maps each word x_i into an embedded input sequence $\boldsymbol{x} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]$, then projects this into a vector $\boldsymbol{h}_i^t \in \mathbb{R}^{d_t}$, where d_t corresponds to the hidden dimension of the obtained text representation. Although our text encoding module is generic, in this paper we implement it as a bi-directional GRU (Cho et al., 2014), on top of pre-trained word embeddings, specifically GloVe (Pennington et al., 2014), as follows.

$$\boldsymbol{h}_i^t = \text{BiGRU}(\boldsymbol{x}_i, \boldsymbol{h}_{i-1}^t) \tag{1}$$

4.2 Audio Encoding Module

We assume the existence of a finite set of time-ordered audio features $\boldsymbol{a} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m]$ extracted from each audio utterance a^s, for instance with the procedure described in Section 5.2. We feed these vectors into another bi-directional GRU to add context to each time step, obtaining hidden states $\boldsymbol{h}_j^a \in \mathbb{R}^{d_a}$.

$$\boldsymbol{h}_j^a = \text{BiGRU}(\boldsymbol{a}_j, \boldsymbol{h}_{j-1}^a) \tag{2}$$

To obtain a condensed representation from the audio signal we again utilize mean pooling over the intermediate memory vectors, obtaining $\bar{\boldsymbol{h}}^a$.

4.3 Video Encoding Module

We propose a video encoding layer that generates a visual representation summarizing spatio-temporal patterns directly from the raw input frames. Concretely, given a video segment $\boldsymbol{v} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_T]$, where $\boldsymbol{v}_i$ is a vector representing a single frame in v^s, our encoding module first maps this sequence into another sequence of video features $\hat{\boldsymbol{v}} = [\hat{\boldsymbol{v}}_1, \ldots, \hat{\boldsymbol{v}}_l]$ following the method described in Section 5.2. Later, this new sequence is mapped into a vector $\bar{\boldsymbol{h}}^v \in \mathbb{R}^{d_v}$ that captures summarized high-level visual semantics in the video, as follows:

$$\boldsymbol{h}_k^v = \text{BiGRU}(\hat{\boldsymbol{v}}_k, \boldsymbol{h}_{k-1}^v) \tag{3}$$

4.4 Fusion Module

We utilize an early fusion strategy similar to Xu et al. (2018) to aggregate the representations obtained from each modality. We concatenate the contextualized representation $\boldsymbol{h}_i^t$ for each token to the summarized representations of the additional modalities, $\bar{\boldsymbol{h}}^a$ and $\bar{\boldsymbol{h}}^v$, and feed this final vector representation to an additional BiGRU:

$$\boldsymbol{h}_i = \text{BiGRU}([\boldsymbol{h}_i^t; \bar{\boldsymbol{h}}^a; \bar{\boldsymbol{h}}^v], \boldsymbol{h}_{i-1}) \tag{4}$$

As a result, our model now allows the representation of each word in the input sentence to interact with the audio and visual features, enabling it to learn potentially different ways to associate each word with the additional modalities. An alternative way to achieve this would be to utilize attention mechanisms to enforce such association behavior, however, we instead let the model learn this relation without using any additional inductive bias.

4.5 Sequence Labeling Module

The main labeling module is a multi-layer perceptron guided by a self attention component. The self attention component enriches the representation $\boldsymbol{h}_i$ with contextual information coming from every other sequence element by performing the following operations:

$$u_{i,j} = \boldsymbol{v}_\alpha^\top \tanh(\boldsymbol{W}_\alpha[\boldsymbol{h}_i; \boldsymbol{h}_j] + \boldsymbol{b}_\alpha) \tag{5}$$

$$\alpha_{i,j} = \text{softmax}(u_{i,j}) \tag{6}$$

$$\boldsymbol{t}_i = \sum_{j=1}^{n} \alpha_{i,j} \cdot \boldsymbol{h}_j \tag{7}$$

$$\boldsymbol{o}_i = \boldsymbol{W}_l[\boldsymbol{h}_i; \boldsymbol{t}_i] + \boldsymbol{b}_l \tag{8}$$

Where $\boldsymbol{o}_i$ is a vector associated to input x_i, and $\boldsymbol{v}_\alpha$, $\boldsymbol{W}_\alpha$, $\boldsymbol{W}_l$ and $\boldsymbol{b}_\alpha$, $\boldsymbol{b}_l$ are trainable parameters. As shown, these vectors are obtained using both the corresponding *aligned* input $\boldsymbol{h}_i$ and the attention-weighted vector $\boldsymbol{t}_i$.

Following previous work, we feed these vectors into a Linear Chain CRF layer, which performs the final labeling. Neural CRFs have proven to be especially effective for various sequence segmentation or labeling tasks in NLP (Ma and Hovy, 2016; Yang and Zhang,

2018; Yang et al., 2018), and have also been used successfully in the past for open domain opinion mining (Zhang et al., 2015). Concretely, we model emission and transition potentials as follows.

$$\psi_i := e(x_i, y_i; \theta) = \boldsymbol{h}_i \cdot \boldsymbol{y}_i \quad (9)$$

$$\psi_{i,j} := q(y_i, y_j; \mathbf{\Pi}) = \mathbf{\Pi}_{y_i, y_j} \quad (10)$$

Where $\boldsymbol{h}_i$ is the fused hidden state for position i and θ denotes the parameters involved in computing this vector, $\boldsymbol{y}_i$ is a one-hot vector associated to y_i, and $\mathbf{\Pi}$ is a trainable matrix of size $\mathbb{L}^{AE}$ or $\mathbb{L}^{C}$ depending on the setting —see Section 5 for more details on this. The score function of a given input sentence s and output sequence of labels l is defined as:

$$\Phi(s, l) = \sum_{i=1}^{n} \log e(x, y_i; \theta) + \log q(y_i, y_{i-1}; \mathbf{\Pi}) \quad (11)$$

In this work we directly optimize the negative log-likelihood associated to this score during training, and apply Viterbi decoding during inference to obtain the most likely labels.

5 Experimental Setup

We evaluate our proposal in several experimental settings based on previous work.

- **Simple**: We only focus on the task of aspect extraction, following a sequence labeling approach with regular IOB tags in $\mathbb{L}^{\text{AE}}$.
- **Collapsed Aspect-Level (CAL)**: We perform aspect extraction and aspect-level sentiment classification with a sequence labeling model, utilizing sentiment-bearing IOB tags in $\mathbb{L}^{\text{C}}$.
- **Collapsed Sentence-Level (CSL)**: Like the previous setting, but we only keep sentence examples that contain a single sentiment, so we can perform sentence-level sentiment classification. Again, we use sequence labeling with sentiment-bearing IOB tags in $\mathbb{L}^{\text{C}}$.
- **Joint Sentence-Level (JSL)**: We use a multi-tasking approach and perform sequence labeling for aspect extraction with regular IOB tags in $\mathbb{L}^{\text{AE}}$, and sequence classification to predict the sentence-level sentiment. In this sense, we add a final 3-layer fully-connected neural network that receives a mean-pooled representation of the fusion layer $\bar{\boldsymbol{h}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{h}_i$ and predicts a sentence-level sentiment. As loss function we utilize the mini-batch average cross-entropy with the gold standard class label. The total loss is the sum of the losses for sequence labeling and sequence classification.

Previous work has also shown that most sentences present a single aspect, and therefore a single sentiment (Marrese-Taylor et al., 2017; Zuo et al., 2018; Zhao et al., 2010), which motivates the introduction of the CSL and JSL settings. For these cases we filtered out sentences that do not fit this description.

5.1 Data

We report results on two different datasets containing fine-grained annotations for both opinion targets and sentiment.

First, we work with the Youtubean dataset (Marrese-Taylor et al., 2017), which contains sentences extracted from YouTube video annotated with aspects and their respective sentiments. The data comes from the user-provided closed-captions derived from 7 different long product review videos about a cell phone, totaling up to 71 minutes of audiovisual data. In total there are 578 long sentences from free spoken descriptions of the product, on average each sentence consist of 20 words. The dataset has a total of 525 aspects, with more than 66% of the sentences containing at least one mention.

Second, we work with the fine-grained annotations gathered for the POM dataset by Garcia et al. (2019b). This dataset is composed of 1000 videos containing reviews where a single speaker in frontal view makes a critique of a movie that he/she has watched. There are videos from 372 unique speakers, with 600 different movie titles being reviewed. Each video has an average length of about 94 seconds and contains 15.1 sentences on average. The fine-grained annotations we utilize are available for each token indicating if it is responsible for the understanding of the polarity of the sentence, and whether it describes the target of an opinion; each sentence has an average of 22.5 tokens. We assume that whenever there is an overlap between the span annotations for a given target and a certain polarity, the corresponding polarity can be assigned to that target, otherwise it is labeled as neutral.

Since the annotated sentences in both datasets are not associated to specific timestamps, in this work we propose a method based on heuristics to rescue the video segments that correspond to each annotated sentence by leveraging video subtitles (or closed-captions.)

```
168
00:20:41,150 —> 00:20:45,109
– How did he do that?
– Made him an offer he could not refuse.
```

Figure 2: Excerpt of a subtitle chunk (in SubRip format,) showing its main components.

As shown in Figure 2, closed captions or subtitles are composed of chunks that contain: (1) A numeric counter identifying each chunk, (2) The time at which the subtitle should appear on the screen followed by −−> and the time when it should disappear, (3) The subtitle text itself on one or more lines, and (4) A blank line containing no text, indicating the end of this subtitle. These chunks exhibit a large variance in terms

of their length, meaning that sentences are usually split into many chunks.

Starting from a subtitle file associated to a given product review video, we apply a fuzzy-matching approach between each annotated sentence for that review and each closed caption chunk. This is repeated for each one of the videos in our datasets. Whenever an annotated sentence matches exactly or has over 90% similarity with a closed caption chunk, its time-span is associated to that sentence. Finally, the "start" and "end" timestamps assigned to each sentence are defined by the start and end time spans of their first and last associated closed captions, sorted by time.

5.2 Implementation Details

Pre-processing for the natural language input is performed utilizing spacy[1], which we use mainly to tokenize. Input sentences are trimmed to a maximum length of 300 tokens, and tokens with frequency lower than 1 are replaced with a special *UNK* marker. To work with the POM dataset, which is already tokenized, we first convert it to the ABSA format, which is tokenization agnostic, and then we process it.

Although our audio encoder is generic, in this work we follow Lakomkin et al. (2017) and use Fast Fourier Transform spectrograms to extract rich vectors from each audio segment. Specifically, we use a window length of 1024 points and 512 points overlap, giving us vectors of size 513. Alternative audio feature extractors such as Degottex et al. (2014) could also be utilized.

On the other hand, in this work we model video feature extraction using I3D (Carreira and Zisserman, 2017). This method inflates the 2D filters of a well-known network e.g. Inception (Szegedy et al., 2015; Ioffe and Szegedy, 2015) or ResNet (He et al., 2016) for image classification to obtain 3D filters, helping us better exploit the spatio-temporal nature of video. We first pre-process the videos by extracting features of size 1024 using I3D with average pooling, taking as input the raw frames of dimension 256×256, at 25 fps. We use the model pre-trained on the kinetics400 dataset (Kay et al., 2017) released by the same authors. Despite our choice to obtain video features, again we note that our video encoder is generic, so other alternatives such as C3D (Tran et al., 2015) could be utilized.

Finally, all of our models are trained in an end-to-end fashion using Adam (Kingma and Ba, 2014) with a learning rate of 10^{-3}. To prevent over-fitting, we add dropout to the text encoding layer. We use a batch size of 8 for the Youtubean dataset, and of 64 for the POM dataset. The language encoder uses a hidden state of size 150, and we fine-tune the pre-trained GloVe.

On each case we compare the performance of our proposed approach against a baseline model that does not consider multi-modality, does not utilize pre-trained GloVe word embeddings and is based on a cross-entropy loss, in which case we simply utilize the mini-batch average cross-entropy between $\hat{\boldsymbol{y}}_i = \text{softmax}(\boldsymbol{o}_i)$ and the gold standard one-hot encoded labels $\boldsymbol{y}_i$, a vector that is the size of the tag label vocabulary for the corresponding task.

[1] https://spacy.io

5.3 Evaluation

Since the size of Youtubean is relatively small, all our experiments in this dataset are evaluated using 5-fold cross validation. In the case of the POM dataset, we report performance on the validation and test sets averaging results for 5 different random seeds. In both cases we compare models using paired two-sided t-tests to check for statistical significance of the differences.

To evaluate our sequence labeling tasks we used the CoNLL *conlleval* script, taking the aspect extraction F1-score as our model selection metric for early stopping. To perform joint aspect extraction and sentiment classification, we considered *positive*, *negative* and *neutral* as sentiment classes, and decoupled the IOB collapsed tags using simple heuristics. Concretely, we recover the aspect extraction F1-score as well as classification performances for each sentiment class.

6 Results

To evaluate the effectiveness of our proposals, we perform several ablation studies on the *Simple* setting for the Youtubean dataset. Using variations of our baseline with pre-trained GLoVe embeddings (GV), conditional random field (CRF), audio and video modalities (A+V). Experiments are also performed using 5-fold cross-validation, and comparisons are always tested for significance using paired two-sided t-tests.

As Table 4 shows, although every proposed model variation performs better than the baseline, only the model uses video and audio modalities obtains a statistically superior performance. We also see that our proposed multi-modal variation is the one that obtains the best performance, also being statistically significant at the highest level of confidence. We believe these results show that our proposed multi-modal architecture is not only able to exploit the features in the audio and video inputs, but it can also leverage the information in the pre-trained word embeddings and benefit from having an inductive bias that is tailored for the task at hand, in this case, with a loss based on structured prediction for sequence labeling.

Table 2 summarizes our results for the Youtubean dataset, where we can see that our proposed multi-modal approach is able to outperform the baseline model for all settings in the aspect extraction task. When it comes to sentiment classification, our multi-modal approaches do not obtain significant performance gain in all cases, sometimes performing worse although without statistical significance. We also compare our results to the performance reported by Marrese-Taylor et al. (2017), who experimented on the *Simple* and *CSL* settings. Their models also use pre-trained word embedding —although different from

Setting	Model	Aspect Extraction			Sentiment Classification		
		P	R	F1	P	R	F1
Simple	Baseline	0.531	0.542	0.533	-	-	-
	Ours	**0.602****	**0.568**	**0.584*****	-	-	-
CAL	Baseline	0.546	0.538	0.539	0.710	0.688	0.696
	Ours	**0.590**	**0.572**	**0.581***	**0.722**	**0.722**	**0.718**
CSL	Baseline	0.526	0.463	0.490	**0.746**	**0.722**	**0.724**
	Ours	**0.563**	**0.581*****	**0.568****	0.720	0.674	0.688
JSL	Baseline	0.483	0.521	0.496	**0.946**	**0.946**	**0.946**
	Ours	**0.544*****	**0.552**	**0.545*****	**0.946**	**0.946**	**0.946**

Table 2: Summary of our results on the Youtubean dataset, *** denotes statistical significance at 99% confidence, ** at 95% and * at 90%.

Setting	Model	Aspect Extraction			Sentiment Classification		
		P	R	F1	P	R	F1
Simple	Baseline	0.394	0.379	0.386	-	-	-
	Ours	**0.396**	**0.406**	**0.399**	-	-	-
CAL	Baseline	0.364	**0.401***	0.382	**0.540*****	0.416	0.270
	Ours	**0.444****	0.368	**0.402****	0.488	**0.466*****	**0.342*****
CSL	Baseline	0.387	0.375	**0.408***	**0.614**	**0.446**	0.296
	Ours	**0.438***	**0.378**	0.404	0.532	**0.446**	**0.304**
JSL	Baseline	0.381	0.357	0.367	0.798	0.802	0.788
	Ours	**0.442*****	**0.401***	**0.420***	**0.924*****	**0.924*****	**0.922*****

Table 3: Summary of our results for the test set of the POM dataset, *** denotes statistical significance at 99% confidence, ** at 95% and * at 90%.

Model	Aspect Extraction		
	P	R	F1
T	0.532	0.543	0.533
T + CRF	0.558	0.528	0.541
T + GV	0.562	0.537	0.548
T + GV + CRF	0.576*	0.569	0.571**
T + A + V	0.587*	0.578	0.580*
T + CRF + A + V	0.578	0.570	0.573*
T + GV + CRF + A + V	0.602**	0.568	0.584***

Table 4: Ablation study on aspect extraction on the simple setting. *** denotes differences against the only text model (T) results are statistically significant at 99% confidence, ** at 95% and * at 90%. (A + V) refers to the audio and video modalities, (GV) stands for GLoVe embeddings and (CRF) for the model trained using the Conditional Random Fields loss.

Setting	Model	AE F1	SC F1
Simple	Baseline	0.428	-
	Ours	**0.433**	-
CAL	Baseline	0.412	0.240
	Ours	**0.427*****	**0.310****
CSL	Baseline	0.408	**0.264**
	Ours	**0.423***	0.262
JSL	Baseline	0.387	**0.950*****
	Ours	**0.469****	0.840

Table 5: Results for the validation set of the POM dataset, where *** denotes results are statistically significant at 99% confidence, ** at 95% and * at 90%.

GloVe— and as input they additionally receives binary features derived from POS tags and other word-level cues. We note, however, that they only experimented with a maximum length of 200 tokens, which makes our results not directly comparable. Their performance on aspect extraction for the *Simple* and *CAL* tasks are 0.561 and 0.555 F1-Score respectively, both of which are lower than ours. In terms of sentiment classification, they report results for each sentiment class with F1-Scores of 0.523, 0.149 and 0.811 for the positive, negative and neutral classes, respectively. Our model is able to outperform this baseline, with a cross-class average F1-Score of 0.718. We do not deepen the analysis in this regard, as numbers are difficult to interpret without statistical testing.

Table 5 and Table 3 summarize our results for the *POM* dataset for the validation and test splits respectively. Compared to the previous dataset we see similar results where our multi-modal approach consistently outperforms the baseline for aspect extraction, but with the gains being comparatively smaller. We also see that our model is able to significantly outperform the baseline in the sentiment classification tasks at least in two

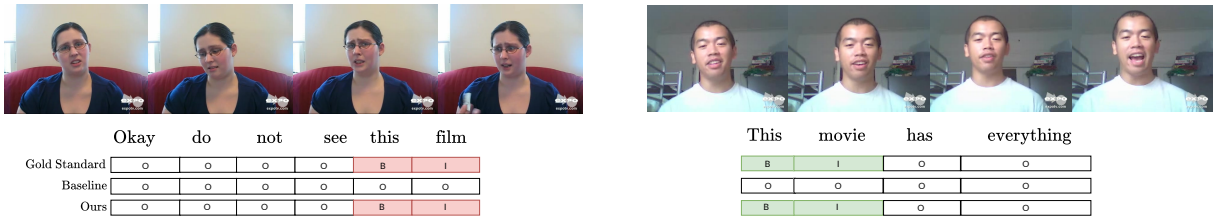

Figure 3: Qualitative comparison between baseline and our method on the POM dataset. Green and red boxes represent positive and negative sentiment respectively.

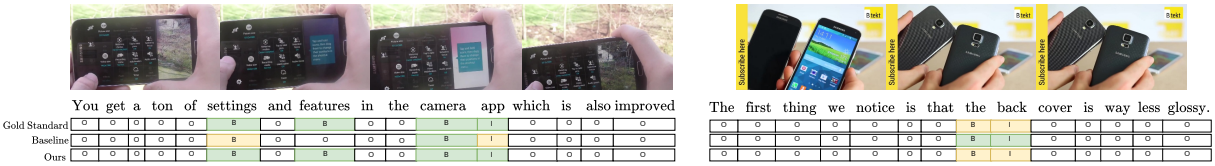

Figure 4: Qualitative comparison between baseline and our method on the Youtubean dataset. Green and yellow boxes represent positive and neutral sentiment respectively.

of out the three settings. In terms of previous work, our results cannot be directly compared to Garcia et al. (2019a) and Garcia et al. (2019b) as their problem setting is different from ours.

On a more broad perspective, we think the performance differences across datasets are related to the nature of each dataset. Meanwhile Youtubean contains reviews about actual physical products, which are often shown in the videos at the same time the reviewer is speaking, the POM dataset contains movie reviews where the speakers directly face the camera during most of the video, without utilizing any additional support material. As a result, the video reviews in the Youtubean dataset mainly focus on capturing images of the products under discussion, with relatively fewer scenes showing the reviewer. This means that there may be few visual cues in the manner of facial expressions or other specific actions that the models could exploit in order to perform better at the sentiment classification task, but more cues useful for aspect extraction. This situation is reverted in the POM dataset, which could explain why our models tend to perform better for sentiment classification, but offering smaller gains for the AE task.

We also think performance differences across datasets are to some extent explained by the nature of the annotations on each case. The annotation guidelines utilized to elaborate each dataset are actually quite different, with the annotations in the Youtubean dataset closely following those of the well-known SemEval datasets, which are target-centric and the POM standards substantially diverging from this. Concretely, Garcia et al. (2019b) propose a two-level annotation method, where "the smallest span of words that contains all the words necessary for the recognition of an opinion" are to be annotated. As a result, aspects annotated in the POM dataset often include pronouns which are more difficult to identify as aspects, often requiring co-reference resolution. With regards to aspect polarity, while it can be extracted directly from the Youtubean annotations, in the case of POM we needed some pre-processing as target and sentiment are annotated using independent text spans.

Qualitative results of the POM and Youtubean dataset in a multitask CAL can be seen in Figure 3 and 4 respectively, results suggest that the method learn to use the information from additional modalities and enhance the sentiment and aspect prediction.

Finally, as we observe that our models tend to obtain bigger gains on the AE tasks rather than on SC, we think this behavior can be partially attributed to the inductive bias of our model, which makes it specially suitable for sequence segmentation tasks.

7 Conclusions

In this paper we have presented a multi-modal approach for fine-grained opinion mining, introducing a modular architecture that utilizes features derived from the audio, video frames and language transcription of video reviews to perform aspect extraction and sentiment classification at the sentence level. To test our proposals we have taken two datasets built upon video review transcriptions containing fine-grained opinions, and introduced a technique that leverages the video subtitles to associate timestamps to each annotated sentence. Our results offer empirical evidence showing that the additional modalities contain useful information that can be exploited by our models to offer increased performance for both aspect extraction and sentiment classification, consistently outperforming text-only baselines.

For future work, we are interested in exploring other ways to capture cross-modal interactions, exploit the temporal relationship between the representations of different modalities, and test alternative ways to better deal with our multi-task settings.

Acknowledgments

We are grateful for the support provided by the NVIDIA Corporation, donating two of the GPUs used for this research.

References

Md Shad Akhtar, Dushyant Chauhan, Deepanway Ghosal, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Multi-task Learning for Multi-modal Emotion Recognition and Sentiment Analysis. In *Proceedings of the 2019 Conference of the North*, pages 370–379, Minneapolis, Minnesota. Association for Computational Linguistics.

Lisa Beinborn, Teresa Botschen, and Iryna Gurevych. Multimodal Grounding for Language Processing. page 15.

Nathaniel Blanchard, Daniel Moreira, Aparna Bharati, and Walter Scheirer. 2018. Getting the subtext without the text: Scalable multimodal sentiment classification from visual and acoustic modalities. In *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*, pages 1–10, Melbourne, Australia. Association for Computational Linguistics.

Darin Brezeale and Diane Cook. 2006. Using closed captions and visual features to classify movies by genre. In *Proceedings of the 7th International Workshop on Multimedia Data Mining (MDM/KDD06): Poster Session*, Washington, DC, USA. ACM.

Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to Answer Open-Domain Questions. pages 1870–1879.

Xinlei Chen, Alan Ritter, Abhinav Gupta, and Tom Mitchell. 2015. Sense Discovery via Co-Clustering on Images and Text. pages 5298–5306.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN EncoderDecoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. Covarepa collaborative voice analysis repository for speech technologies. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 960–964. IEEE.

Desmond Elliott, Stella Frank, Loc Barrault, Fethi Bougares, and Lucia Specia. 2017. Findings of the Second Shared Task on Multimodal Machine Translation and Multilingual Image Description. In *Proceedings of the Second Conference on Machine Translation*, pages 215–233, Copenhagen, Denmark. Association for Computational Linguistics.

Alexandre Garcia, Pierre Colombo, Florence d'Alché-Buc, Slim Essid, and Chloé Clavel. 2019a. From the Token to the Review: A Hierarchical Multimodal approach to Opinion Mining. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5542–5551, Hong Kong, China. Association for Computational Linguistics.

Alexandre Garcia, Slim Essid, Florence d'Alch Buc, and Chlo Clavel. 2019b. A multimodal movie review corpus for fine-grained opinion mining. *arXiv:1902.10102 [cs]*. ArXiv: 1902.10102.

Deepanway Ghosal, Md Shad Akhtar, Dushyant Chauhan, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. 2018. Contextual Inter-modal Attention for Multi-modal Sentiment Analysis. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3454–3466, Brussels, Belgium. Association for Computational Linguistics.

S. Gupta and R.J. Mooney. 2009. Using closed captions to train activity recognizers that improve video retrieval. In *Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. IEEE Computer Society Conference on*, pages 30–37.

Sonal Gupta and Raymond J. Mooney. 2010. Using closed captions as supervision for video activity recognition. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI-2010)*, pages 1083–1088, Atlanta, GA.

Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed (Ehsan) Hoque. 2019. UR-FUNNY: A multimodal language dataset for understanding humor. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2046–2056, Hong Kong, China. Association for Computational Linguistics.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

Ozan Irsoy and Claire Cardie. 2014. Opinion Mining with Deep Recurrent Neural Networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 720–728, Doha, Qatar. Association for Computational Linguistics.

Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. 2017. The kinetics human action video dataset. *CoRR*.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*.

Amar Krishna, Joseph Zambreno, and Sandeep Krishnan. 2013. Polarity trend analysis of public sentiment on youtube. In *Proceedings of the 19th International Conference on Management of Data*, COMAD '13, pages 125–128, Mumbai, India, India. Computer Society of India.

Egor Lakomkin, Cornelius Weber, and Stefan Wermter. 2017. Automatically augmenting an emotion dataset improves classification using audio. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 194–197, Valencia, Spain. Association for Computational Linguistics.

Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2015. Combining Language and Vision with a Multimodal Skip-gram Model. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 153–163, Denver, Colorado. Association for Computational Linguistics.

Pengfei Liu, Shafiq Joty, and Helen Meng. 2015. Fine-grained Opinion Mining with Recurrent Neural Networks and Word Embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1433–1443, Lisbon, Portugal. Association for Computational Linguistics.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.

Edison Marrese-Taylor, Jorge Balazs, and Yutaka Matsuo. 2017. Mining fine-grained opinions on closed captions of YouTube videos with an attention-RNN. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 102–111, Copenhagen, Denmark. Association for Computational Linguistics.

Grgoire Mesnil, Xiaodong He, Li Deng, and Yoshua Bengio. 2013. Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding. In *INTERSPEECH*, pages 3771–3775.

Margaret Mitchell, Jacqui Aguilar, Theresa Wilson, and Benjamin Van Durme. 2013. Open Domain Targeted Sentiment. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1643–1654, Seattle, Washington, USA. Association for Computational Linguistics.

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*, ICMI '14, pages 50–57, New York, NY, USA. ACM.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*.

Hai Pham, Thomas Manzini, Paul Pu Liang, and Barnabs Poczs. 2018. Seq2seq2sentiment: Multimodal Sequence to Sequence Models for Sentiment Analysis. In *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*, pages 53–63, Melbourne, Australia. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 Task 4: Aspect Based Sentiment Analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

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 the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2539–2544, Lisbon, Portugal. Association for Computational Linguistics.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-Dependent Sentiment Analysis in User-Generated Videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–883, Vancouver, Canada. Association for Computational Linguistics.

Aliaksei Severyn, Alessandro Moschitti, Olga Uryupina, Barbara Plank, and Katja Filippova. 2014. Opinion Mining on YouTube. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1252–1261, Baltimore, Maryland. Association for Computational Linguistics.

Stefan Siersdorfer, Sergiu Chelaru, Wolfgang Nejdl, and Jose San Pedro. 2010. How useful are your comments?: Analyzing and predicting youtube comments and comment ratings. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 891–900, New York, NY, USA. ACM.

Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. 2016. A Shared Task on Multimodal Machine Translation and Crosslingual Image Description. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 543–553, Berlin, Germany. Association for Computational Linguistics.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.

Yasuyuki Tahara, Atsushi Tago, Hiroyuki Nakagawa, and Akihiko Ohsuga. 2010. Nicoscene: Video scene search by keywords based on social annotation. In Aijun An, Pawan Lingras, Sheila Petty, and Runhe Huang, editors, *Active Media Technology*, volume 6335 of *Lecture Notes in Computer Science*, pages 461–474. Springer Berlin Heidelberg.

Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. 2016a. Effective LSTMs for Target-Dependent Sentiment Classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3298–3307, Osaka, Japan. The COLING 2016 Organizing Committee.

Duyu Tang, Bing Qin, and Ting Liu. 2016b. Aspect Level Sentiment Classification with Deep Memory Network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 214–224, Austin, Texas. Association for Computational Linguistics.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497.

Bin Wu, Erheng Zhong, Ben Tan, Andrew Horner, and Qiang Yang. 2014. Crowdsourced time-sync video tagging using temporal and personalized topic modeling. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 721–730, New York, NY, USA. ACM.

Huijuan Xu, Kun He, Bryan A. Plummer, Leonid Sigal, Stan Sclaroff, and Kate Saenko. 2018. Multilevel Language and Vision Integration for Text-to-Clip Retrieval. *arXiv:1804.05113 [cs]*. ArXiv: 1804.05113.

Liheng Xu, Kang Liu, and Jun Zhao. 2014. Joint Opinion Relation Detection Using One-Class Deep Neural Network. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 677–687, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Jie Yang, Shuailong Liang, and Yue Zhang. 2018. Design Challenges and Misconceptions in Neural Sequence Labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3879–3889, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Jie Yang and Yue Zhang. 2018. NCRF++: An Open-source Neural Sequence Labeling Toolkit. In *Proceedings of ACL 2018, System Demonstrations*, pages 74–79, Melbourne, Australia. Association for Computational Linguistics.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor Fusion Network for Multimodal Sentiment Analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114, Copenhagen, Denmark. Association for Computational Linguistics.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. MOSI: Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos. *arXiv:1606.06259 [cs]*. ArXiv: 1606.06259.

Meishan Zhang, Yue Zhang, and Duy Tin Vo. 2015. Neural Networks for Open Domain Targeted Sentiment. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 612–621, Lisbon, Portugal. Association for Computational Linguistics.

Xin Zhao, Jing Jiang, Hongfei Yan, and Xiaoming Li. 2010. Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 56–65, Cambridge, MA. Association for Computational Linguistics.

Y. Zuo, J. Wu, H. Zhang, D. Wang, and K. Xu. 2018. Complementary aspect-based opinion mining. *IEEE Transactions on Knowledge and Data Engineering*, 30(2):249–262.