INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Opinion Dynamics Modeling for Movie Review Transcripts Classification with Hidden Conditional Random Fields

Valentin Barriere[1], *Chloé Clavel*[1], *Slim Essid*[1]

[1]LTCI, Télécom ParisTech, Université Paris Saclay, F-75013, Paris, France
{firstname.lastname}@telecom-paristech.fr

Abstract

In this paper, the main goal is to detect a movie reviewer's opinion using hidden conditional random fields. This model allows us to capture the dynamics of the reviewer's opinion in the transcripts of long unsegmented audio reviews that are analyzed by our system. High level linguistic features are computed at the level of inter-pausal segments. The features include syntactic features, a statistical word embedding model and subjectivity lexicons. The proposed system is evaluated on the ICT-MMMO corpus. We obtain a F1-score of 82%, which is better than logistic regression and recurrent neural network approaches. We also offer a discussion that sheds some light on the capacity of our system to adapt the word embedding model learned from general written texts data to spoken movie reviews and thus model the dynamics of the opinion.

Index Terms: Hidden Conditional Random Field, Opinion Mining, Linguistic Patterns, Word Embedding

1. Introduction

With the growing importance of social networks, the amount of Internet user data has increased dramatically in the last few years. It is now important for companies to exploit this new source of information about their customers in order to be more competitive. The concept of some websites is even to be simply a huge database of recommendations, such as *rottentomatoes.com* where the users rate and review movies, thus delivering their opinion about those movies.

The domain of opinion mining in textual documents has developed considerably in the last several years. The trend is to use deep learning approaches that allow for achieving high performance, relying on a big amount of training labeled data [21]. On the other hand, hybrid approaches [26] combine the robustness and the high accuracy of Machine Learning (ML) algorithms with the fine-grained modeling of linguistic rules. They do not require a huge amount of labeled data and thus are an interesting alternative to deep learning methods.

As far as the representation of the data is concerned, various alternatives have been considered in previous works in this area. Using the negations and the intensifiers present in the context of the word as input features for a machine learning algorithm has initially been studied by [8] on the textual IMdB movie review database. The Bag-of-Words (BOW) is a classical domain-agnostic paragraph representation. [15] used BoW and SVM for a sentiment analysis task over the MOUD dataset (Vlogs) obtaining a score of 64.94%. The Bag-of-N-grams (BoNG), which is an extension of this model, was used by [19] for a sentiment analysis task over the Metacritic database (textual movie reviews). [23] merge the results of subjectivity lexicons, valence shifters and BoNG to train a classifier for sentiment analysis in tweets. Another trendy option to represent the data nowadays is to create a distributed vector of every word in an unsupervised way, training the model on a large dataset of text. In [16], authors use word2vec with a CNN-SVM for a binary valence classification task on short speech utterances while in [7] the authors use the same representation for a task of subjective expression extraction from sentences. In this paper we are in line with all these studies since we combine distributional word embedding with lexicons, linguistic patterns, syntactic features and paralinguistic cues to train a learning model.

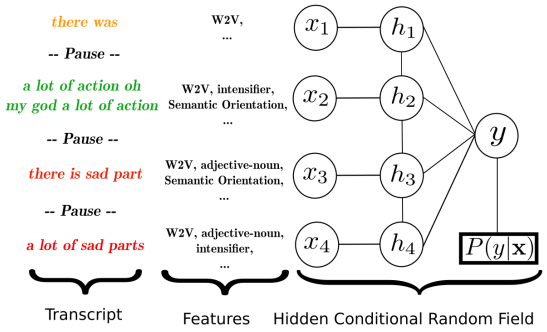
Another issue that is tackled in this paper is how to deal with opinion dynamics in long reviews where the speaker develops his/her opinion across the review. For example, a negative review can include some expressions of positive opinion and then end by a negative opinion. When the size of the documents is increasing, it is crucial to account for the dynamics of the document by using relevant ML method. Opinion dynamics modeling has been rarely addressed in the literature of opinion mining. While some studies are restricted to real-time prediction by segmenting the document into several parts of fixed duration [17], others insist on the complementarity of the modalities in order to detect multimodal patterns [3,13]. We distinguish ourselves from these studies by focusing the analysis on the textual modality while using the audio modality only to segment the text based on the pauses of the reviewer. This is motivated by the idea of modeling the opinion dynamics in a more natural way.

Thus, we investigate a latent state model in order to model the opinion of a speaker along a globally annotated audio movie review. The absence of written punctuation prevent us to segment using syntax so we choose to use oral pauses because of the relevant role of those self-interruptions in the segmentation of discourses [4]. Here, we consider the task of labeling an audio transcript with respect to opinions using a variant of Conditional Random Fields (CRF), a discriminative classifier that has proven its utility in several NLP and Computer Vision tasks. This variant, called Hidden Conditional Random Fields (HCRF) has been successfully used to analyze sequences of textual, audio or visual to be labeled globally with only one output [18]. Latent state models have already proven their efficiency for multimodal sentiment analysis or agreement classification [3, 13]. The objective here is to investigate the potential of HCRF for a classification using transcripts from oral speech. The discriminative nature of CRF will enable some strong linguistic rules combined with other features to emerge directly from the learning phase.

In the second section of this paper, we will present the features we chose for our task and our learning model. In the third section, we will present the dataset, talk about our experiments and results and finish in the fourth section with a discussion of the results and then we will conclude our paper.

http://dx.doi.org/10.21437/Interspeech.2017-1035

Figure 1: *Overview of the system*



2. Feature and classification model description

2.1. Overview of the system

Because of the structure of spontaneous speech, a lot of sentences are unfinished, making it difficult to segment a spoken review into relevant units. We choose to use the pauses to segment the review into Inter Pausal Units (IPUs). Then, we produce the features for each IPU and use them to feed the HCRF, which predicts the most probable label for the current review (see figure 1).

2.2. Features

We can sort the textual features we use into 4 groups :
- *The N-grams features* : The BoNG presented in [19] is an extension of the classical Bag of Words representation to N-grams. In this work we use words, bi-grams and tri-grams.
- *The distributed representations* : word2vec is a distributed learning model to represent words [10]. The principle is to use the surrounding words to find the general context in which a word appears and learn its weights statistically. During the learning phase, the vectors of the words appearing in the same context are expected to get closer. This representation can be used to learn more specific semantic information about the discourse of the speaker in the textual features. We chose to use word2vec since it has been found to give better results on a sentiment analysis task in [16] compared to other statistical word embeddings. The 300-dimensional vectors we used were pretrained over a corpus of 100 billions words from Google Press[1]. Generally, it has been empirically found that a more general and bigger training dataset makes it possible to obtain vectors that perform better on several tasks [11].
- *The linguistic and lexicon-based features* : The affective valence of a document can be directly retrieved with a rule-based heuristic using specific values attributed for each word with lexicons. We use the negative, positive and neutral SentiWordNet (SWN) scores [1] and the dominance, arousal and valence scores of the enriched Affective norms for English Words (ANEW) lexicon [24]. We use linguistic patterns such as adjectives followed by a noun, negations, intensifiers (amplifiers and downtoners). We decided to combine linguistic patterns with sentiment lexicons using the Semantic-Orientation CALculator (SO-CAL) [22] which is composed of a grouping of lexicons containing subjective words, intensifiers and valence shifters with associated values. Those values are used for arithmetic operations following simple patterns to give a semantic orientation score to a sentence (see details [22]). We separate each value into 3 scores reflecting a positive, a negative and a neutral score so that they can be independently significant of different emotional states of the speaker. We finally take the disfluencies, the presence of a capital letter and the 6 parts-of-speech from [16] plus interjections and pronouns which are significant of emotional bursts, or belongings.
- *The Paralinguistic features* : The paralinguistic information provided in the transcript can indicate an emotional state which the reviewer does not necessarily evoke through words. The 8 main paralinguistic annotations were dispatched in different categories : the intonation, the pronunciation, the laughter and the volume.

[1]Details at https://code.google.com/archive/p/word2vec/

2.3. Classification Model

The HCRF model is used in order to learn a mapping from a sequence of observations $\mathbf{x}_i = \{x_1, ..., x_{L_i}\}$ of length L_i to a label $y_i \in \mathcal{Y}$. Each observation x_k is represented by a feature vector $\phi(x_k)$. For every $\mathbf{x}_i$, a sequence of unobserved latent variables $\mathbf{h}_i = \{h_1, ..., h_{L_i}\}$ is defined where $h_k \in \mathcal{H}$, $\mathcal{H}$ being a finite set of states [18].
The label decision is made using the posterior probability $P(y|\mathbf{x}, \theta)$ given by Eq (1), where θ refers to the parameters of the HCRF.

$$P(y|\mathbf{x}, \theta) = \sum_{\mathbf{h}} P(y, \mathbf{h}|\mathbf{x}, \theta) = \frac{\sum_{\mathbf{h}} e^{\Psi(y, \mathbf{h}, \mathbf{x}; \theta)}}{\sum_{y', \mathbf{h}} e^{\Psi(y', \mathbf{h}, \mathbf{x}; \theta)}}, \quad (1)$$

where $\Psi(y, \mathbf{h}, \mathbf{x}; \theta) \in \mathbb{R}$ is a potential function (defined in Eq (2)) that measures the compatibility between a label, a sequence of hidden states and the observations. The definition depends on different types of feature functions described below:

$$\Psi(y, \mathbf{h}, \mathbf{x}; \theta) = \sum_j \langle \phi(x_j) \,|\, \theta_o(h_j) \rangle + \sum_j \theta_s(y, h_j) + \sum_j \theta_t(y, h_j, h_{j+1}) \quad (2)$$

- *The hidden state feature functions* depend only on the current observation vector and the current hidden state. A weight $\theta_o(h_j)$ is created for each hidden state h_j. The inner product represents the compatibility between an observation and the hidden state.
- *The label feature functions* depend on the label and the current state. The weight $\theta_s(y, h_j)$ represents the compatibility between a label y and a hidden state h_j.
- *The hidden state transition feature functions* depend on the position in the sequence and the label. The weight $\theta_t(y, h_j, h_{j+1})$ represents the compatibility between a label y and the transition from a hidden state h_j to an other hidden state h_{j+1}.
The model is classically trained by minimizing an ℓ_2-norm regularized negative log-likelihood cost [18]. Decision is taken by choosing the label y that maximizes $P(y|\mathbf{x}, \theta)$.

3. Experiments and results

We tested three models with different feature sets and segmentations in order to validate our models. Firstly, we created a baseline for our task using a logistic regression model with Bag-of-N-Gram features at document level. Since logistic regression does not take into account the dynamics of the observations, we tried a more powerful alternative baseline model that can handle sequential data: a recurrent neural network (RNN-LSTM) [6].

Compared to these models, HCRF offers the benefit of interpretability in the way it handles sequential data, while having the potential to model the dynamics of opinion-related phenomena (emotional states, stances, etc.) through latent states.

Moreover, we compared BoNG to our feature set and tested different pause-based segmentations.

We used a 10-fold Cross-Validation (CV) where train and test sets are disjoint to validate our models with each test part containing the same proportion of both classes as in the total dataset.

3.1. Dataset

In this study, we used the ICT-MMMO corpus[2] consisting of 365 movie review videos obtained from Youtube.com and ExpoTV.com [25]. Those reviews are performed by non-professional users and the audio quality of the recording varies significantly. All the videos of the corpus have been annotated in valence by one or two independent annotators. The valence score goes from 1, which means that the speaker has a very negative opinion about the movie, to 5 which denotes a strongly positive opinion about the movie from the speaker, and 3 meaning neutral. The reference is obtained by taking the mean of scores given by the two annotators on a video. The dataset contains more positive videos than negative videos (opinion annotations of the videos : 120 negative, 38 neutral, 207 positive). All the video clips are manually transcribed to extract the spoken words. Using the Transcriber software [2] each spoken utterance is segmented according to the pause duration. All the annotations, the transcriptions of the text and paralinguistic events were made without using the visual information.

We decided to discard the neutral files because they include files annotated with a different polarity by the two annotators. We obtained a total of 321 videos (116 negative and 205 positive) for a total of 13h12 of audio, composed of **12625** segmented IPUs and **143181** words.

3.2. Baselines using LogReg and LSTM

Methodology : We considered a baseline model with a simple textual feature set and with our feature set that we tested for different textual representation levels (at the document level or using the pauses) in order to measure the improvement brought by the HCRF. We used logistic regression with a BoNG model like [25] with the same parametrization: applying trigram features, Porter stemming, TF-IDF transformations, and document-length normalization. We kept a larger vocabulary. We then changed for a more sophisticated feature set (our set in Table 1), that is a representation using the statistical word embedding model from [11] described in 2.2. After a tokenization[3] we used a spell checker[4] to eliminate the numerous typos from the transcription and to clean the text before taking the word-vectors (stop-words excluded). We followed the protocol of [11] addressing a sentiment analysis task on short texts and we aggregated by averaging the representations of every word contained inside the IPU to obtain one vector of the same size, and standardized them. In order to help the determination of the opinion we added the linguistic rule set and the values of the subjectivity lexicons (as described in 2.2). We used the number of linguistic patterns we detected as well as the scores from the subjectivity lexicons for every word on each IPU to obtain one score per feature on each IPU. We standardized each linguistic feature. We used pauses longer than 150, 300 and 500 ms (3 experiments) to segment the documents into IPUs.Regarding the tuning of the logistic regression hyperparameters, we trained with values of the inverse of the regularization strength C in $\{0.1, 0.5, 1, 10, 100\}$. We used the scikit-learn [14] implementation of logistic regression. For the RNN-LSTM, we used the keras implementation [5] with a number of hidden states in $\{64, 128, 256\}$, a dropout regularization of U and W (see [6]) in $\{0.1, 0.2, 0.3\}$ (higher dropout decreased performances) and a number of epochs in $\{4...10\}$. We used the cross entropy as cost function and Adam as learning algorithm [9].

Results : The results of the baselines are listed in the first part of Table 1 using F1-scores and accuracy. In this table, the global $F1$ (the harmonic mean of recall and precision) is the average $F1$ of both classes ($F1+$ and $F1-$) weighted by their priors and $Accuracy$ is the percentage of true predictions. We notice that the best results are obtained with our feature set. This result is actually unexpected given that we are averaging all the word-vectors of the document into a single one, but the effectiveness comes from the other sentiment-related and linguistic features. The results of the RNN-LSTM are not better for the negative class. Though it has the potential to capture some dynamics, the neural network requires more data than available in the considered corpus to be fully effective. Using the BoNG baseline, [19] obtained a F1-score of 78.74% for a sentiment analysis task over the Metacritic database (textual movie reviews).

3.3. HCRF models

Methodology : The existence of latent states in HCRF makes them useful to model a dynamic system like, for example, the emotional state of the speaker. Using our feature set, including sentiment-related features and a distributed representation, the model is expected to more effectively exploit the concepts employed by the speakers. We also investigated the granularity of the segmentation, using different thresholds to use the pauses to segment. We trained the HCRF model with the Matlab wrapper of the HCRF Library [12] and used a L-BFGS solver for the training. Regarding the exploration of the model hyperparameters, we trained with different values for : the ℓ_2 regularization parameter in $\{0.01, 0.05, 0.075, 0.1, 0.25, 0.5, 1\}$, the context window in $\{0, 1, 2\}$ and the number of hidden states in $\{2...5\}$. The context window is the number of IPU neighbors we concatenated with the centered IPU. We also tested more hidden states, without better results but a longer training time.

Results : The results with the HCRF models are summarized in Table 1. The best configuration was obtained with 5 hidden states, no context and a value of the regularization parameter equal to 10. As expected, the HCRF improves the results compared to the logistic regression (F1 score improves from 79 to 80) with the BoNG features. The best results were obtained using our set with an improvement of the negative class F1 score (8 points). The 300-ms threshold also brings a slight improvement to the negative class, while using a higher threshold (500 ms) decreases the performance. However, a 10-fold CV does not bring enough information to conclude about the statistical significance of this difference in performance ($p = 0.15$ for the negative class).

4. Discussion

False predictions : We provide here an in-depth analysis of false predictions. We found that many examples of wrong clas-

[2]data available by sending an email to abagherz@cs.cmu.edu

[3]We used the CoreNLP from Standford [20]

[4]https://github.com/phatpiglet/autocorrect

Table 1: *F1-scores and accuracies results with different feature sets, segmentation thresholds and models*

Features	Model	F1+	F1-	F1	Acc
Majority label	Dummy	78	0	50	63
BoNG	LogReg	84	69	78	79
Our set	LogReg	83	72	79	79
Our set (150ms)	LSTM	84	68	78	78
BoNG (150ms)	HCRF	84	67	78	79
Our set (150ms)	HCRF	85	72	80	80
Our set (300ms)	HCRF	**86**	**75**	**82**	**82**
Our set (500ms)	HCRF	82	67	77	77

sification were due to an opinion too briefly expressed in a review which was globally neutral about the movie. The corresponding videos contain too few linguistic cues of the global opinion of the review. The system also seems to be influenced by the portions of the review where the speaker relates other people's opinion or where they express a strong opinion about something or somebody. There are also cases where the speaker briefly leaves his/her opinions at the beginning or at the end of the video and the main part of the review consists of the reviewer's opinions about general things that are not concerning neither the movie nor its features. Thus, the prediction is complex. The algorithm did not have enough examples to check that the most important points are the opinions of the speaker related to the movie and its features. Finally, when examining the pause segmentation ground-truth, we see some segmentation errors: there are 108 IPUs containing more than 50 words (using the 150-ms threshold). Besides, errors are dispatched over more than 18% of the files of the corpus. A clean pause detection method based on a text aligner could be an effective solution to this problem.

Hidden states, transitions and activation words : After each training of a HCRF model, there is, for each label y at least one state h_y with compatibility weight $\theta_s(y, h_y)$ that is highly positive with one label and highly negative with the other label. The transition between those states is highly improbable. We will call those states '*negative state*' (*Neg*) and '*positive state*' (*Pos*) even if it is an abuse of language. The three other states are considered 'neutral' (*Neu1*, *Neu2*, *Neu3*), with low amplitude transition and compatibility weights. Those states can be used as a bridge between positive and negative states to model the development of the opinion dynamics. In Table 2, we present the most relevant examples of the most compatible features with each hidden state (features that correspond to the 30 highest positive weights). In the first column, we can see that the linguistic and paralinguistic features have a less important weight in the neutral states: the only feature having a positive weight for all the neutral states is '**chuckling**' while *Pos* and *Neg* have numerous and various linguistic and paralinguistic features with high positive weights.

Regarding word embedding features, our system is no longer learning words but concepts in the 300-dimensional word2vec space by using the information contained inside the word-vectors. In order to analyze the features of the word2vec space, we look for vectors of the words contained in our corpus that activated each state the most. In the second column of Table 2, we can see activation words with high valences, e.g. '*disappointing*', '*miserably*' and '*awesome*'.

Role of neutral states : Learning word embeddings requires a significant amount of text data to be available, that is the reason why we choose to use pre-trained word embeddings. It is interesting to notice that, even though the used word-vectors were learned from general text data, they include spontaneous speech words, such as '*uhm*' or '*dunno*'. However, they do not correspond to the ones that would have been learned on an audio monologue such as the reviews analyzed in this work. For example, while a written '*uhm*' in a post may be a stylistic effect aiming at sounding negative, the oral counterpart is a common hesitation and is possibly neutral. Another example is the difference between *yes* and *yeah*: the latter is not common in written text where it reflects a more positive thought (see Table 3). Further, some other words are merely corpus-specific, e.g. *Hi* and *Thanks* (*Thanks for watching me guys*) but associated with positive valence by their word2vec trained on text data. Consequently, information inside the word-vectors may sometimes not be adapted to the discourse of the speaker. The hidden neutral states of the HCRF seem to be handling this issue, so that the problematic word-vectors do not affect the states linked with the global labels of the review.

Table 2: *Most relevant examples of features with high positive values for each state (paralinguistic with *)*

States	Linguistic and paralinguistic features	Words corresponding to compatible vectors
Pos	adj, disfluency, conjunction, intensifier, *lip smacking*, ...	*honors*, *fearless*, *awesome*, *fantastic*
Neu1	*chuckling*	*um*, *Uh*, *ah*, *dunno*, *nada*
Neu2	*chuckling*	*um*, *Uh*, *ah*, *dunno*, *nada*
Neu3	∅	*Thanks*, *justin*, *sean*, *michael*, *Sorry*
Neg	negation, *falling intonation*, interjection, *word elongation*, ...	*miserably*, *disappointing*, *yelling*, *failure*, *lack*

Table 3: *Examples of differences in feature function values*

Words	Positive State	Neutral States	Negative State
Uh	-2.57	2.8	-3.86
Yeah	0.98	2.21	-5.49
Yes	-0.06	1.21	-3.14
Thanks	2.25	3.48	-7.41

5. Conclusion and future work

In this paper, we have presented a HCRF model that uses a pause-based segmentation of movie review transcripts in order to model the dynamics of the opinion of the speaker through latent states. Our textual feature set includes word embedding, linguistic rules and clues from subjectivity lexicon. The use of HCRF classifiers allows us to implicitly learn local linguistic representations of each inter-pausal segment of the reviews making the integration of word embeddings in the classification system more meaningful. We also investigated a pause-based segmentation on a long unannotated discourse, finding that too long segments lead to a loss of performance.

In our future work we would like to improve the way we use the word embedding in our model in order to handle more precise concepts with more hidden states. Further, we would like to test on a bigger corpus in order to obtain significant results.

6. References

[1] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 0(January):2200–2204, 2010.

[2] Claude Barras, Edouard Geoffrois, Zhibiao Wu, and Mark Liberman. Transcriber: Development and use of a tool for assisting speech corpora production. *Speech Communication*, 33(1-2):5–22, 2001.

[3] Konstantinos Bousmalis, Louis-Philippe Morency, and Maja Pantic. Modeling hidden dynamics of multimodal cues for spontaneous agreement and disagreement recognition. In *2011 IEEE International Conference on Automatic Face and Gesture Recognition and Workshops, FG 2011*, pages 746–752, 2011.

[4] Estelle Campione and Jean Véronis. A large-scale multilingual study of pause duration. In *Speech Prosody 2002. Proceedings of the1st International Conference on Speech Prosody*, pages 199–202, 2002.

[5] Francois Chollet. Keras, 2015.

[6] Sepp Hochreiter and J Urgen Schmidhuber. LONG SHORT-TERM MEMORY. *Neural Computation*, 9(8):1735–1780, 1997.

[7] Ozan Irsoy and Claire Cardie. Opinion mining with deep recurrent neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 720–728, 2014.

[8] Alistair Kennedy and Diana Inkpen. Sentiment classification of movie reviews using contextual valence shifters. In *Computational Intelligence*, volume 22, pages 110–125, 2006.

[9] Diederik Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations*, pages 1–13, 2014.

[10] Tomas Mikolov, Greg Corrado, Kai Chen, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. *Proceedings of the International Conference on Learning Representations (ICLR 2013)*, pages 1–12, 2013.

[11] Tomas Mikolov, I. Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. In *Proc. NIPS*, pages 1–9, 2013.

[12] Louis-Philippe Morency. Hidden-state Conditional Random Field (HCRF) Library, 2007.

[13] Louis-philippe Morency, Rada Mihalcea, and Payal Doshi. Towards Multimodal Sentiment Analysis: Harvesting Opinions from the Web. *Proceedings of the 13th International Conference on Multimodal Interfaces (ICMI-11)*, pages 169–176, 2011.

[14] Fabian Pedregosa, Gal Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and douard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2012.

[15] Veronica Perez-Rosas, Rada Mihalcea, and Louis-Philippe Morency. Utterance-Level Multimodal Sentiment Analysis. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 973–982, 2013.

[16] Soujanya Poria, Erik Cambria, and Alexander Gelbukh. 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*, page 25392544, 2015.

[17] Soujanya Poria, Erik Cambria, Newton Howard, Guang Bin Huang, and Amir Hussain. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing*, 174:50–59, 2016.

[18] Ariadna Quattoni, Sybor Wang, Louis-Philippe Morency, Michael Collins, and Trevor Darrell. Hidden conditional random fields. *IEEE transactions on pattern analysis and machine intelligence*, 29(10):1848–1853, 2007.

[19] Bjorn Schuller, Joachim Schenk, Gerhard Rigoll, and Tobias Knaup. "The Godfather" vs. "Chaos": Comparing linguistic analysis based on on-line knowledge sources and bags-of-N-grams for movie review valence estimation. *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, pages 858–862, 2009.

[20] Sebastian Schuster and Christopher D. Manning. Enhanced English Universal Dependencies: An Improved Representation for Natural Language Understanding Tasks. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2371–2378, 2016.

[21] Richard Socher, Alex Perelygin, and Jy Wu. Recursive deep models for semantic compositionality over a sentiment treebank. *EMNLP-2013: Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, 2013.

[22] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*, 37(2):267–307, 2011.

[23] Joachim Wagner, Piyush Arora, Santiago Cortes, Utsab Barman, Dasha Bogdanova, Jennifer Foster, and Lamia Tounsi. DCU: Aspect-based Polarity Classification for SemEval Task 4. *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, (SemEval):223–229, 2014.

[24] Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, 45 VN - r(4):1191–1207, 2013.

[25] Martin Wöllmer, Felix Weninger, Tobias Knaup, Bjorn Schuller, Congkai Sun, Kenji Sagae, and Louis-Philippe Morency. YouTube Movie Reviews: Sentiment Analysis in an Audio-Visual Context. *IEEE Intelligent Systems*, 28(3):46–53, 2013.

[26] Bishan Yang and Claire Cardie. Joint Inference for Fine-grained Opinion Extraction. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics,*, pages 1640–1649, 2013.