World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

Sentiment analysis using Hierarchical Multimodal Fusion (HMF)

Bishwo Prakash Pokharel * and Roshan Koju

Faculty of Science, Health and Technology, Nepal Open University, Bagmati, Nepal.

World Journal of Advanced Research and Reviews, 2022, 14(03), 296–303

Publication history: Received on 08 May 2022; revised on 12 June 2022; accepted on 14 June 2022

Article DOI: https://doi.org/10.30574/wjarr.2022.14.3.0549

Abstract

The rapid rise of platforms like YouTube and Facebook is due to the spread of tablets, smartphones, and other electronic devices. Massive volumes of data are collected every second on such a platform, demanding large-scale data processing. Because these data come in a variety of modalities, including text, audio, and video, sentiment categorization in various modalities and emotional computing are the most researched fields in today's scenario. Companies are striving to make use of this information by developing automated systems for a variety of purposes, such as automated customer feedback collection from user assessments, where the underlying challenge is to mine user sentiment connected to a specific product or service. The use of efficient and effective sentiment analysis tools is required to solve such a complex problem with such a big volume of data. The sentiment analysis of videos is investigated in this study, with data available in three modalities: audio, video, and text. In today's world, modality fusion is a major problem. This study introduces a novel approach to speaker-independent fusion: utilizing deep learning to fuse in a hierarchical fashion. The work tried to obtain improvement over simple concatenation-based fusion.

Keywords: Multi-modal; Bimodal; Sentiment analysis; Hierarchical fusion; Emotion

1. Introduction

Internet is easily available to all people since the dawn of this millennium. As a result, numerous social media sites, such as YouTube, Facebook, Instagram, and others, have sprung up where people can express their varied viewpoints on a variety of topics using various approaches such as postings, photographs, and videos. The fast development of electronic devices has dramatically accelerated the distribution of content in numerous ways. People have been giving their opinions on various materials, products, events, services, and other topics in various forms since lately. This content is one of the most important sources for large corporations to examine evaluations in multidimensional ways by extracting user sentiment, ideas, and complaints from video reviews.

Videos carry information by employing three channels: audio, video, and a transcript of speech. Sentiment analysis in many modalities is opinion mining from these several modalities of data, and the main issue is fusing the various features (1).

1.1. Sentiment Analysis

The main goal is to assign the correct sentiment polarity to a given text, which could be a sentence, utterance, or document. Positive polarity, negative polarity, and neural polarity are the three major polarities that are considered.

* Corresponding author: Bishwo Prakash Pokharel
Faculty of Science, Health and Technology, Nepal Open University, Bagmati, Nepal.

1.2. Sentiment analysis with multiple modalities

Videos often have three different information channels: an audio transcript, audio itself, and video. Modalities are another name for these channels. As a result, assigning sentiment polarity to such a film requires sentiment analysis in several modalities. Text is more likely than other modalities to contain sentiment information. Visual modality, on the other hand, may provide additional information not caught by textual modality. Facial expressions and small muscular movements are examples of such information (such as frown, grin etc.). The audio modality contributes to the sentiment classification process by providing crucial information such as pitch, tone shift, and so on.

2. Related Works

Text-based sentiment analysis approaches are divided into two categories: knowledge-based methods and statistics-based methods (2). These strategies have a wide range of approaches and their importance changes throughout time. In the past, knowledge-based systems were more widely utilized, but statistics-based systems have recently dominated tremendously; particularly, supervised statistical methods have played a large role.

In terms of other modalities, substantial research in the early 1970s revealed that facial expressions provide sufficient indications for emotion identification (3). Recent research has begun to focus on acoustic characteristics such as pitch, utterance intensity, bandwidth, and duration (4).

Early works (5)(6), We can see that combining audio and visual to create a bimodal signal resulted in higher precision than single-mode systems. Analysis of such fusion was done at both feature level (7) and decision level (8).

Furthermore, substantial work has been done on fusing audio and video for emotion recognition, as well as uncovering the role of text in multimodal emotion detection mechanisms alongside audio and video. Few works (9)(10) is done to fuse information from different attributes like audio, video and text extracting the real emotion and sentiment. Fusion of two modalities, audio and text, has also been done for emotion recognition (11)(12). Both methods were based on feature-level fusion. Audio and text have also been combined at the decision-making level (13). In the current case, (14) use CNN for feature extraction, followed by multiple-kernel learning (MKL) for emotion recognition and sentiment analysis.

3. Methodology

3.1. Block Diagram

The feature vectors of three modalities are first renovated to have the same dimensions in this study. These modified vectors are thought to contain abstract features that represent the properties relevant to sentiment categorization. Then, using fully-connected layers of a neural network, each of these abstract properties is compared and combined. As a result, fusion bimodal feature vectors are produced. Finally, the bimodal vectors are transformed into a trimodal vector using fully connected layers. The comparison and analysis are based on the features generated in this manner, and the best for the sentiment classification task is recommended.

Extraction of text, audio, and video from their sources, pre-processing extracted data, feeding information to respective base models, training base models, and fusing the output from base models and training, providing final prediction are all part of the multi-modal sentiment analysis implementation process.

3.2. Dataset Used

The majority of sentiment analysis research involving several modalities is conducted on datasets with two splits, namely, train and test, which may share specific speakers. Because each person expresses their feelings and sentiments in their own unique way, it's critical to identify general, person-independent characteristics.

3.2.1. CMU-MOSI

The CMU-MOSI dataset(15) contains a wealth of sentiment expressions, with 89 people reviewing diverse topics in English. Five annotators divide videos into distinct utterance levels with ratings ranging from +3 (highly positive) to -3 (strongly negative). The sentiment polarity is calculated using the average of these five annotations, and only two classes are evaluated (positive and negative).

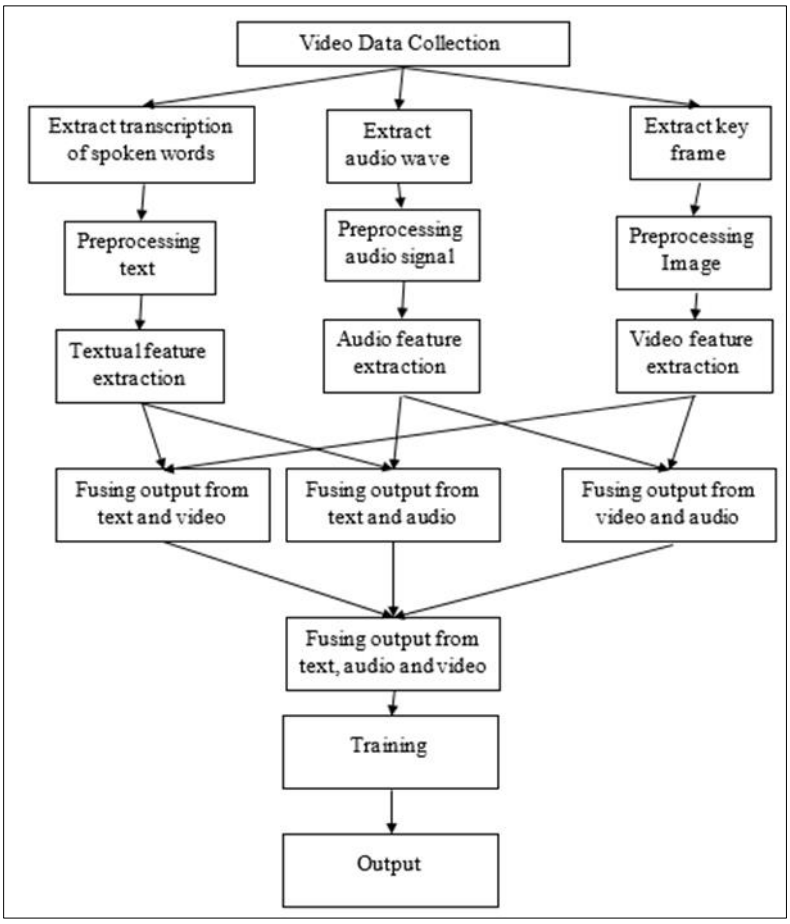

Figure 1 Building blocks of fusing text, audio and video

3.3. Feature extraction

3.3.1. Extracting textual features

The transcription of people's spoken utterances is the most common source of textual modality. CNN (a deep convolutional neural network)(16) is used to extract features from textual data. First, each utterance of constituent words is represented by a concatenation of vectors, which in our tests contains 100 billion words collected from publicly available Google News with 300-dimensional word2vec vectors trainings(17). Null vectors of exactly 50 words are also used for truncation or padding of utterances.

3.3.2. Extracting feature from audio

The audio features are extracted at a frame rate of 30Hz with a sliding window of 100ms. An open source software named openSMILE(18) is used to compute the features. Audio attributes were used to automatically extract pitch and voice intensity. Furthermore, samples are recognized by normalizing the voice and thresholding the voice strength to distinguish between samples with and without voice. When it comes to voice normalization, the Z-standardization method is used. OpenSMILE is in charge of both of these tasks.

3.3.3. Extracting feature from video

The 3D-CNN approach is used to extract visual information from the video. It is hypothesized that 3D-CNN will be able to learn not just relevant information from each frame, but also changes over a specified number of consecutive frames. T3D-CNN has been successfully used to classify objects in 3D data(19). We chose to employ it because of its capacity to produce cutting-edge results.

3.4. Hierarchical Multimodal Fusion

The mechanism for fusing unimodal features vectors is:

- Acoustic feature ($f_a \in \mathbb{R}^{da}$)
- Visual feature ($f_v \in \mathbb{R}^{dv}$)
- Textual feature ($f_t \in \mathbb{R}^{dt}$)

The unimodal features may have different dimensionalities, i.e., $d_a \neq d_v \neq d_t$. So, it is mapped them to the same dimensionality, say d_m, using fully-connected layer as follows:

- $g_a = \tanh (W_a f_a + b_a)$, (1)
- $g_v = \tanh (W_v f_v + b_v)$, (2)
- $g_t = \tanh (W_t f_t + b_t)$, (3)

Where, $W_a \in \mathbb{R}^{dm \times da}$, $b_a \in \mathbb{R}^{dm}$, $W_v \in \mathbb{R}^{dm \times dv}$, $b_v \in \mathbb{R}^{dm}$, $W_t \in \mathbb{R}^{dm \times dt}$ *and* $b_t \in \mathbb{R}^{dm}$, we represent the mapping for each dimension as:

$$g_x = (c_1^x, c_2^x, c_3^x, \ldots, c_{dm}^x) \quad (4)$$

Where $x \in \{v, a, t\}$ and c_l^x are scalars for all $l = 1, 2, \ldots, d_m$. We can see these values c_l^x as more abstract than that of derived value from fundamental feature values (which are the components of f_a, f_v, and f_t). For example, an abstract feature can be the angriness of a speaker in a video. It is inferred the degree of angriness from visual features (f_v; facial muscle movements), acoustic features (f_a; pitch, raised voice etc.), or textual feature (f_t; the language, choice of words etc.). Therefore, the degree of angriness can be represented by c_l^x where x = a, v. t and *l* are some fixed integers between 1 and d_m.

Now, the assessment of nonconcrete feature values from different modalities discussed here may have contradiction with each other. So, for making comparison among the feature values network is essential. To achieve combination of two modalities (which are audio-video, audio-text, and video-text) is taken at a time and compared-and-combined their feature values (i.e., c_l^v with c_l^t, c_l^v with c_l^a, and c_l^a with c_l^t) respectively using fully-connected layers as follows:

$$i_l^{va} = \tanh (w_l^{va}.[c_l^v, c_l^a]^T + b_l^{va}), \quad (5)$$

$$i_l^{at} = \tanh (w_l^{at}.[c_l^a, c_l^t]^T + b_l^{at}), \quad (6)$$

$$i_l^{vt} = \tanh (w_l^{vt}.[c_l^v, c_l^t]^T + b_l^{vt}), \quad (7)$$

where $w_l^{va} \in \mathbb{R}^2$, b_l^{va} is scalar, $w_l^{at} \in \mathbb{R}^2$, b_l^{at} is scalar, $w_l^{vt} \in \mathbb{R}^2$, b_l^{vt} is scalar, for $l = 1, 2, \ldots, d_m$. A hypothesis is made on that it will enable network for comparison of the decisions from each modality against the others and help achieve a better fusion of modalities.

3.4.1. Bimodal fusion

The bimodal fused features for video-audio, audio-text, video-text are defined as:

$$F^{va} = (i_1^{va}, i_2^{va}, \ldots, i_{dm}^{va}), \quad (8)$$

$$F^{at} = (i_1^{at}, i_2^{at}, \ldots, i_{dm}^{at}), \quad (9)$$

$$F^{vt} = (i_1^{vt}, i_2^{vt}, \ldots, i_{dm}^{vt}), \quad (10)$$

respectively.

3.4.2. Trimodal fusion

All three modalities are combined using fully-connected layers as follows:

$$Z_l = \tanh (w_l.[i_l^{va}, i_l^{at}, i_l^{vt}]^T + b_l), \quad (11)$$

where $w_l \in \mathbb{R}^3$, and b_t is a scalar for all $l = 1, 2, \ldots, d_m$. So, we define the fused features as:

$$F_{avt} = (z_0, z_1, \ldots, z_{dm}). \quad (12)$$

3.5. Classification

In order to perform classification, the fused features *F* are fed to a layer called SoftMax with *C = 2* outputs. The classifier can be described as follows:

$$P = \text{SoftMax}\ (W_{softmax}\ F^{q} + b_{softmax}) \quad (13)$$

$$\hat{y} = \ argmax_j(p[j]) \quad (14)$$

Where, $W_{softmax} \in \mathbb{R}^{c \times dm}$, $b_{softmax} \in \mathbb{R}^{C}$, $P \in \mathbb{R}^{C}$, j = class value (0 or 1), q = va, at, vt, avt, and $\hat{y}$ = estimated class value.

3.6. Training

Categorical cross-entropy as loss function (J) is employed for training,

$$J = \ -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=0}^{C-1} y_{ij} log P_i[j] \quad (15)$$

Where N = number of samples

i = index of a sample,

j = class value, and

$$y_{ij} = \begin{cases} 1, & if\ expected\ class\ value\ of\ sample\ i\ is\ j \\ 0, & otherwise \end{cases}$$

The network for 200 epochs is trained for the bimodal and 100 epochs are trained for the trimodal features, where optimization of the parameter set θ = {W_a, W_v, W_t, b_a, b_v, b_t, w_1^{va}, w_2^{va}, …, w_{dm}^{va}, w_1^{at}, w_2^{at}, …, w_{dm}^{at}, w_1^{vt}, w_2^{vt}, …, w_{dm}^{vt}, b_1^{va}, b_2^{va}, …, b_{dm}^{va}, b_1^{at}, b_2^{at}, …, b_{dm}^{at}, b_1^{vt}, b_2^{vt}, …, b_{dm}^{vt}, w_1, w_2, …, w_{dm}, b_1, b_2, …, b_{dm}, $W_{softmax}$, $b_{softmax}$}is done.

4. Results and discussion

The studies were carried out using Python version 3 in the Spyder framework, with Keras 2.0 or higher, TensorFlow 1.7 or higher, NumPy, and Scikit-learn.

The confusion matrix and classification report are used in this study to assess the model's performance. Precision, recall, and F1 score are all included in the classification report.

$$Precision = \frac{True\ Positive}{(True\ Positive + False\ Positive)} \quad (16)$$

$$Recall = \frac{True\ Positive}{(True\ Positive + False\ Negative)} \quad (17)$$

$$F1\ Score = 2 * \frac{(Precision * Recall)}{(Presicion + Recall)} \quad (18)$$

- A true positive show that the class was successfully predicted.
- A false positive is when a class is predicted mistakenly when it is not that class.
- Predicted incorrectly since the other class is False Negative.

4.1. Experiments

The results of our HMF model are compared to the three baselines we created (1), (20) and (21). Textual features are extracted using a modal with a Convolutional neural network, while visual and audio features are extracted using the CLM-Z and OpenSMILE toolkits, respectively, in Poria et al.(21). The bimodal and trimodal feature vectors are created by concatenating unimodal features. Zadeh et al.(20) present a new feature fusion paradigm termed tensor fusion. Majumder et al.(1) employ a multimodal fusion method.

Table 1 Comparison of different paper

Modality	Poria et al. (21)	Majumder et al. (1)	Zadeh et al.(20)	Our HMF
T+A	73.2	74.2	77.0	80.6
T+V	73.2	74.4	77.1	85.5
A+V	55.7	57.5	56.5	75.1
T+A+V	73.5	74.6	77.2	80.0

The three basis publications that used the identical CMU-MOSI dataset are compared to our HMF on a broad scale. In comparison to baselines, our HMF model performs better in terms of classification accuracy in all modalities, as shown in the table. It has been observed that our method's performance accuracy has improved, and it now displays the optimal outcome in all combinations.

A significant indicator is found among the abstract feature values after examining the data. Our strategy produces greater outcomes when using text with video and audio with video, indicating that video is the most powerful individual modality. We may deduce that the people's true emotions are mirrored in their facial expressions, as evidenced by the video.

Poria et al.(21) can only fuse using concatenation in their method. Our HMF method outperforms their approach by a wide margin. The approach of Zadeh et al.(20), on the other hand, uses Tensor to fuse the data. Our HMF approach is increasingly outperforming them, thanks to bimodal fusion and the possibility of using it for trimodal fusion in the future.

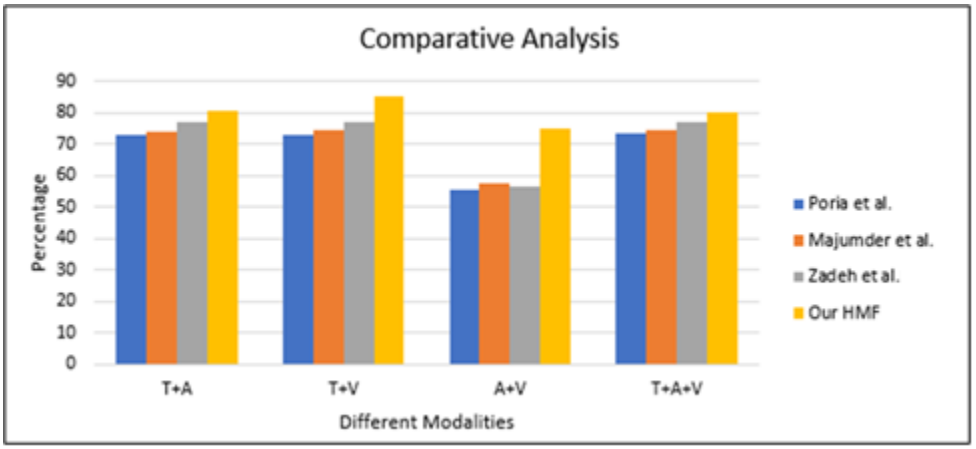

Figure 2 Comparison of experimented results on CMU-MOSI dataset

When compared to earlier modalities of three base papers, the result of our HMF appears to be superior. This is the consequence of the test and train dataset's disjoint pattern. The speaker independent experiment is carried out for sentiment analysis of several modalities in order to improve the results, which is the paper's unique addition.

5. Conclusion

The multimodal fusion technique is an important aspect of sentiment analysis in several modalities. The key issues in multimodal sentiment analysis are emotion recognition, contextual information extraction, and multimodal fusion. The person-independent hierarchical multimodal strategies for feature extraction using the LSTM model are presented in this paper. In this paper, an analytical and comprehensive fusion strategy is proposed. Our solution outperformed the commonly used early fusion on the CMU-MOSI dataset, which was developed to test multimodal sentiment analysis methods.

It is expected to improve the quality of unimodal characteristics, particularly textual features, in the future, which would boost classification accuracy. The impact of a subcategory of features and class-specific features on classification accuracy is also being investigated in the future.

Compliance with ethical standards

Acknowledgments

This research work is done by Bishwo Prakash Pokharel under the supervision of Dr. Roshan Koju, under the Faculty of Science, Health and Technology, Nepal Open University, Nepal.

Disclosure of conflict of interest

The authors declare no conflict of interest.

References

[1] Majumder N, Hazarika D, Gelbukh A, Cambria E, Poria S. Multimodal sentiment analysis using hierarchical fusion with context modeling. Knowledge-Based Syst [Internet]. 2018;161(October 2017):124–33. Available from: https://doi.org/10.1016/j.knosys.2018.07.041

[2] Cambria E. Affective Computing and Sentiment Analysis. IEEE Intell Syst. 2016;31(2):102–7.

[3] Tivatansakul S, Ohkura M, Puangpontip S, Achalakul T. Emotional healthcare system: Emotion detection by facial expressions using Japanese database. 2014 6th Comput Sci Electron Eng Conf CEEC 2014 - Conf Proc. 2014;41–6.

[4] Datcu D, Rothkrantz LJM. Semantic Audiovisual Data Fusion for Automatic Emotion Recognition. Emot Recognit A Pattern Anal Approach. 2015;411–35.

[5] De Silva LC, Miyasato T, Nakatsu R. Facial emotion recognition using multi-modal information. Proc Int Conf Information, Commun Signal Process ICICS. 1997;1(October):397–401.

[6] Chen LS, Huang TS, Miyasato T, Nakatsu R. Multimodal human emotion/expression recognition. In: Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition. IEEE Comput. Soc; 2002

[7] Kessous L, Castellano G, Caridakis G. Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis. J Multimodal User Interfaces. 2010;3(1):33–48.

[8] Schuller B. Recognizing affect from linguistic information in 3D continuous space. IEEE Trans Affect Comput. 2011;2(4):192–205.

[9] Wollmer M, Weninger F, Knaup T, Schuller B, Sun C, Sagae K, et al. You tube movie reviews: Sentiment analysis in an audio-visual context. IEEE Intell Syst. 2013;28(3):46–53.

[10] Rozgić V, Ananthakrishnan S, Saleem S, Kumar R, Prasad R. Ensemble of SVM trees for multimodal emotion recognition. In: Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference. 2012. p. 1–4.

[11] Metallinou A, Lee S, Narayanan S. Audio-visual emotion recognition using Gaussian mixture models for face and voice. In: 2008 Tenth IEEE International Symposium on Multimedia. IEEE; 2008.

[12] Eyben F, Wöllmer M, Graves A, Schuller B, Douglas-Cowie E, Cowie R. On-line emotion recognition in a 3-D activation-valence-time continuum using acoustic and linguistic cues. J Multimodal User Interfaces. 2010;3(1):7–19.

[13] Wu CH, Liang W Bin. Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels (Extended abstract). 2015 Int Conf Affect Comput Intell Interact ACII 2015. 2015;31(4):477–83.

[14] Poria S, Cambria E, Gelbukh A. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. Conf Proc - EMNLP 2015 Conf Empir Methods Nat Lang Process. 2015;(September):2539–44.

[15] Zadeh A, Zellers R, Pincus E, Morency LP. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. IEEE Intell Syst. 2016;31(6):82–8.

[16] Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Li FF. Large-scale video classification with convolutional neural networks. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit. 2014;1725–32.

[17] Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. 1st Int Conf Learn Represent ICLR 2013 - Work Track Proc. 2013;1–12.

[18] Eyben F, Schuller B. openSMILE:). ACM SIGMultimedia Rec. 2015;6(4):4–13.

[19] Ji S, Xu W, Yang M, Yu K. 3D Convolutional neural networks for human action recognition. IEEE Trans Pattern Anal Mach Intell. 2013;35(1):221–31.

[20] Zadeh A, Chen M, Cambria E, Poria S, Morency LP. Tensor fusion network for multimodal sentiment analysis. EMNLP 2017 - Conf Empir Methods Nat Lang Process Proc. 2017;1103–14.

[21] Poria S, Chaturvedi I, Cambria E, Hussain A. Convolutional MKL based multimodal emotion recognition and sentiment analysis. Proc - IEEE Int Conf Data Mining, ICDM. 2017;439–48.