TF-MVSA: Multimodal Video Sentiment Analysis Tool using Transfer Learning

Victor Akihito Kamada Tomita
Universidade de São Paulo
São Carlos, Brasil
akihito012@usp.br

Ricardo Marcondes Marcacini
Universidade de São Paulo
São Carlos, Brasil
ricardo.marcacini@icmc.usp.br

Abstract

Existing methods for sentiment analysis in videos rely on extensive training on large labeled datasets, making them expensive and impractical for real-world applications. This challenge becomes even more complex when dealing with labeled data in different modalities. To address these limitations, we proposed a transfer learning method and a computational tool that leverage pre-trained models for each modality and employ modality consensus to automatically annotate video segments. Our tool implements neural networks with attention mechanisms to learn the significance of each modality during the learning process. The experimental results demonstrate that our tool surpasses unimodal methods and remains competitive with multimodal approaches, even when labeled data for analyzing new videos are unavailable. Moreover, the tool is publicly available, thereby serving as a competitive baseline for similar multimodal sentiment analysis methods.

Keywords: video sentiment analysis, transfer learning, multimodal learning

1 Introduction

The ever-expanding landscape of various media and social networks, including platforms like YouTube and news sites, led to a vast proliferation of videos expressing opinions on diverse subjects [4]. This surge in video content prompted companies to seek ways to extract valuable insights from these videos, enabling them to comprehend their audience's perceptions of their products. Additionally, political parties have a vested interest in understanding public opinions, particularly concerning the population's voting intentions [3].

In response, sentiment analysis aims to automatically extract specific emotional reactions expressed by entities towards particular objects [11]. Text-based sentiment analysis systems have been developed to focus primarily on a speech from interlocutors [9], including captions and audio transcriptions while neglecting other crucial characteristics like visual attributes. Relying solely on speech transcription overlooks valuable data for evaluation, such as gestures and facial expressions [2]. Consequently, there has been a recent surge in multimodal sentiment analysis [5], which explores diverse video data representations by establishing connections between unimodal models.

Our focus lies on developing a multimodal model with the capability to analyze both visual and textual characteristics of a video, extracting the expressed sentiment over time. While existing studies have presented promising solutions for multimodal video sentiment analysis [11], there are drawbacks that limit their applicability in real-world scenarios. The latest studies predominantly rely on deep learning methods, necessitating large datasets for model training. Thus, we address the following question: *"How can we train a multimodal sentiment analysis model in an unsupervised manner?"*

Our proposed method entails two key steps. Firstly, we explore pre-trained general-purpose models specifically designed for sentiment analysis in images (facial expressions) and texts (captions and transcriptions). Leveraging these pre-existing models allows us to extract valuable features from the visual and textual components of the videos. Secondly, we present neural network that incorporates attention mechanisms during model training [7]. This neural network uses the consensus reached from the sentiment classifications obtained through each unimodal model. By integrating the outputs of these models with attention mechanisms, we enhance the multimodal sentiment analysis.

The principal highlight of our proposed method, called Transfer Learning-based Multimodal Video Sentiment Analysis (TF-MVSA), is its ability to conduct sentiment analysis for videos without the need for labeled data during the training stage. We conducted an evaluation of the the impact of each modality (textual and visual) on the fusion layer of the neural network in a controlled transfer learning scenario. Our experimental results showcased that fusing the modalities using the attention mechanism and consensus strategy from our model yielded superior performance compared to individual (unimodal) models for each modality. Additionally, this fusion approach proved to be competitive when compared to (multimodal) voting strategies of the modalities.

In: XXII Workshop de Ferramentas e Aplicações (WFA 2023), Ribeirão Preto, Brasil. Anais Estendidos do Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia). Porto Alegre: Sociedade Brasileira de Computação, 2023.

ISSN 2596-1683

2 Multimodal Sentiment Analysis

Sentiment analysis is an automated method for extracting and analyzing judgments about entities, using diverse representations like enthusiasm, valence, and dominance. The literature also presents discrete approaches for emotions and sentiment representation [6].

Recently, there is a growing interest in multimodal sentiment analysis [12], focusing on exploring visual, textual, and audio features. This has given rise to applications like video sentiment analysis, demonstrated by projects such as MOSI, MELD, and IEMOCAP [4, 8], enabling comprehensive analysis of text, visual content, and audio.

Multimodal models can be classified into three categories: Late fusion, Early fusion, and Hybrid fusion [10]. Late fusion involves training unimodal models separately and merging their outputs using voting mechanisms. Early fusion combines unimodal models at the start of data processing to create a single representation for multimodal training. Hybrid Fusion combines aspects of both late and early fusion methods.

The approach proposed in this paper falls under the category of Hybrid Fusion, since we combine pre-trained models for each modality and employing modality consensus during the annotation of video segments (early fusion), and then we merge the outputs of these unimodal models using attention mechanisms and neural networks (late fusion).

3 Transfer Learning-based Multimodal Video Sentiment Analysis (TF-MVSA)

The architecture of the proposed method, TF-MVSA explores multimodal sentiment analysis using hybrid fusion. It combines emotion recognition from facial expressions and sentiment analysis from captions to classify the sentiment of videos. Figure 1 provides an overview of the TF-MVSA method.

TF-MVSA employs a fine-tuned RoBERTa model for sentiment classification, as proposed by Zhang et al. [13], enabling general-purpose sentiment analysis for new videos. To extract visual features from facial expressions and for sentiment classification in faces, we adopt a real-time optimized convolutional neural network (CNN) introduced by Valdenegro-Toro et al. [1].

The pre-trained models mentioned earlier are utilized to train the multimodal model through unimodal consensus. Once the unimodal analysis is complete, we align the results and train the multimodal model specifically for cases where the unimodalities exhibit agreement. For the cases where the unimodalities do not reach a consensus, we consider the classification output provided by the TF-MVSA multimodal model as the final result.

In our multimodal approach, we employ fusion layers with an attention mechanism. This merging technique aims to train the model to discern the relative importance of each modality concerning sentiment polarity. As a result, our TF-MVSA method enables the adaptation of modality importance on the agreements obtained in the late fusion stage.

One of the key advantages of TF-MVSA is its ability to accomplish sentiment classification without requiring labeled data from the new video. This is achieved through training a multimodal model that takes inputs from the text and image models' embeddings, while utilizing the outputs obtained from pre-trained models as labels. By adopting this transfer learning approach, the model can effectively generalize to new video domains. In brief, the main functionalities of this TF-MVSA tool are:

- **Fusion of Multimodal Information**: By incorporating embeddings from both text and image models as inputs, TF-MVSA achieves multimodal fusion, effectively integrating information from multiple sources. This integration enhances the model's overall performance and comprehension.
- **Generalization to Unseen Domains**: Benefiting from its transfer learning approach, the model demonstrates robust generalization to previously unseen video domains, making it versatile and adaptable to diverse contexts.

4 Experimental Evaluation

To understand the TF-MVSA training performance, we create synthetic embeddings simulating various noise levels and error rates, enabling a controlled transfer learning scenario. We compare the results with simple unimodal models through the training of dense networks and random models. Additionally, we measure the performance difference between the late fusion model (baseline) and our hybrid model.

Figure 2a provides an overview of our experimental results. Each point represents the $F1-score$ obtained by testing a specific model, while the box plots depict the distribution of points associated with its corresponding color. The blue points represent the performance of our proposed multimodal hybrid fusion model, the red points symbolize the unimodal image model, the green points depict the unimodal text model, the purple points characterize the multimodal late fusion model, and the orange points symbolize a random model.

We observed that the $F1 - score$ distribution obtained by the two unimodal models correlates with the total noise embedded in their respective modalities. The multimodal hybrid fusion model achieved the best results, followed by the multimodal late fusion model. This finding demonstrates the effectiveness of our proposed TF-MVSA approach in identifying the significance of each modality during the fusion stage.

Figure 2b illustrates the mean distribution of the $F1-score$ obtained by each model concerning the probability of the

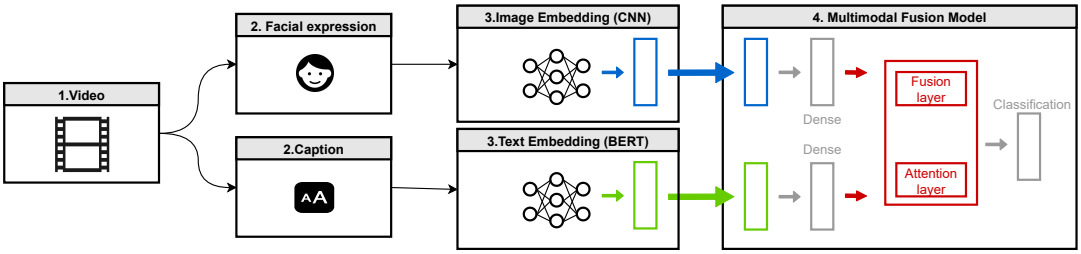


Figure 1. Overview of the proposed TF-MVSA tool.

unimodal image model correctly classifying sentiments. For the unimodal text model, we use the complement of this probability (e.g., if the image model has a 0.75 probability of correct sentiment classification, the text model has a 0.25 probability).

In this figure, the hybrid fusion model's curve resembles a parabola. As the accuracy of the image model increases, the hybrid model prioritizes the image modality, and when the image model's accuracy is low, it prioritizes the textual modality. However, when the probabilities of both modalities are close to 50%, the multimodal hybrid model decides which modality to prioritize, leading to decreased performance. It is important to highlight the significant improvement TF-MVSA achieved compared to the traditional late fusion model. Additionally, multimodal models demonstrate their effectiveness compared to unimodal models.

4.1 TF-MVSA tool

Based on the analyses conducted earlier, a software tool for sentiment analysis for YouTube videos has been developed. Figure 3 demonstrates an example of the TF-MVSA tool in action. The three images present respectively, sentiment analysis from caption text, sentiment analysis from facial expression. multimodal sentiment analysis using the hybrid fusion approach. The tool combines the strengths of unimodal models to enhance sentiment analysis accuracy and provide valuable insights from different modalities.

From a social perspective, this computational tool holds significant potential for social organizations. By utilizing TF-MVSA to analyze sentiments expressed in videos available on social media, these organizations can gain valuable insights into public sentiment towards various social causes. From a business perspective, companies can leverage this tool to gain valuable insights into what their users truly think and feel about their products and services. Finally, from an academic perspective, we believe that TF-MVSA is as a competitive baseline for multimodal sentiment analysis methods. Its ability to leverage existing pre-trained models and perform sentiment analysis without relying on annotated data in the target videos is a significant advantage.

4.2 Privacy, License, and Access

This academic tool is a technologically advanced and scientifically developed solution, available for private, commercial, or legal entities use without territorial restrictions. It operates under an open-source GNU General Public License, allowing users to freely use, modify, and share the software. The tool does not retain ownership of the data collected, nor does it bear responsibility for its usage.

5 Concluding Remarks

The paper introduces unsupervised multimodal video sentiment analysis using a transfer learning strategy with pre-trained models. The proposed method generates the TF-MVSA tool, that utilizes a hybrid fusion strategy to combine textual and visual embeddings from these models. The evaluation transfer learning over synthetic datasets and a demo tool applicable to YouTube videos.

We provide a demonstration video showcasing TF-MVSA's output and the source code for result reproducibility:

- Demonstration video: TF-MVSA's output on a benchmark video from the MOSI repository. Available at: https://youtu.be/VWjdd9-3CNs
- TF-MVSA Source code: An anonymized GitHub repository with installation and execution instructions. Additionally, a Jupyter Notebook guides TF-MVSA execution from a YouTube video ID. Available at: https://github.com/Vakihito/SentimentYoutube

While our Transfer Learning-based Multimodal Video Sentiment Analysis (TF-MVSA) tool shows promising results, there are some directions for future research:

- **Real-time Analysis**: Optimizing TF-MVSA for real-time sentiment analysis to enable applications in live video streams and social media platforms.
- **Multilingual Support**: Extending TF-MVSA to support sentiment analysis in videos with diverse languages to cater to a wider range of content.
- **User Interaction**: Incorporating user feedback mechanisms, such as active learning, to continuously improve the TF-MVSA tool performance over time.

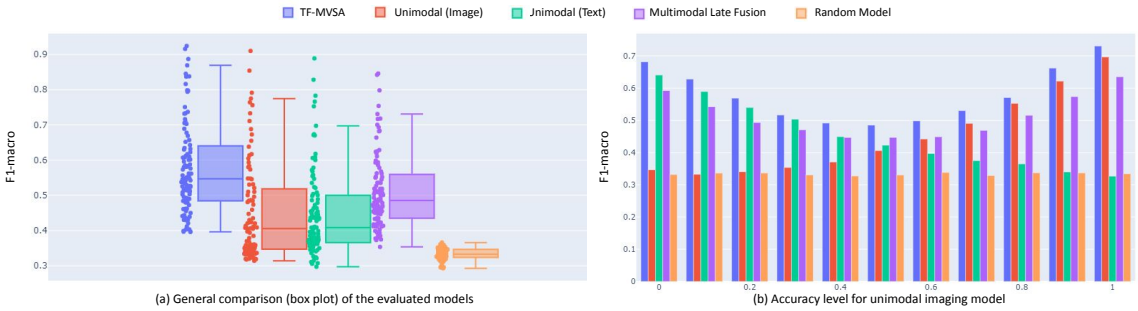
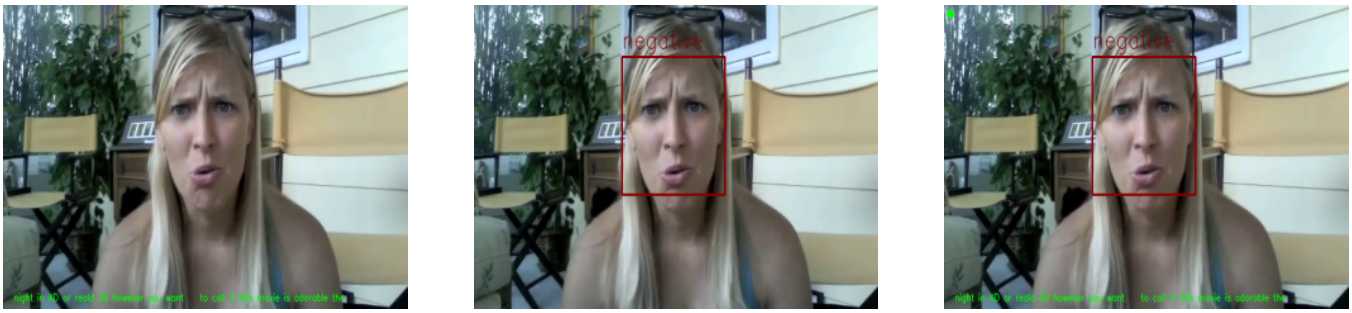


Figure 2. Comparison of experimental results.



(a) Text Caption (Unimodal) **(b)** Facial Expression (Unimodal) **(c)** TF-MVSA (Multimodal)

Figure 3. TF-MVSA tool demonstration over a MOSI benchmark video. In the textual modality, sentiments are color-coded in captions (green=positive, red=negative, white=neutral). For the visual modality, sentiment labels appear above face rectangles. In the multimodal version, a consensus sentiment is displayed in the top-left corner, combining both modalities' information.

Acknowledgments

This work was supported by CAPES and CNPq. The authors of this work would like to thank the Center for Artificial Intelligence (C4AI-USP) and the support from the FAPESP grant 2019/07665-4 and from the IBM Corporation. Finally, this work was supported by *Ministério da Ciência, Tecnologia e Inovações*, with funds from Law 8248, of October 23, 1991, PPI-SOFTEX, coordinated by Softex [DOU 01245.010222/2022-44].

References

[1] Octavio Arriaga, Matias Valdenegro-Toro, and Paul Plöger. 2019. Real-time Convolutional Neural Networks for emotion and gender classification. In *27th European Symposium on Artificial Neural Networks, ESANN 2019, Bruges, Belgium, April 24-26, 2019*. 221–226.

[2] Ringki Das and Thoudam Doren Singh. 2023. Multimodal sentiment analysis: A survey of methods, trends and challenges. *Comput. Surveys* (2023).

[3] Bernard J Jansen, Mimi Zhang, Kate Sobel, and Abdur Chowdury. 2009. Twitter power: Tweets as electronic word of mouth. *Journal of the American society for information science and technology* 60, 11 (2009).

[4] Taeyong Kim and Bowon Lee. 2020. Multi-attention multimodal sentiment analysis. In *International Conference on Multimedia Retrieval*. 436–441.

[5] Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. 2011. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th international conference on multimodal interfaces*. 169–176.

[6] Myriam Munezero, Calkin Suero Montero, Erkki Sutinen, and John Pajunen. 2014. Are they different? Affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE transactions on affective computing* 5, 2 (2014), 101–111.

[7] Zhaoyang Niu, Guoqiang Zhong, and Hui Yu. 2021. A review on the attention mechanism of deep learning. *Neurocomputing* 452 (2021), 48–62.

[8] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508* (2018).

[9] Annamaria Porreca, Francesca Scozzari, and Marta Di Nicola. 2020. Using text mining and sentiment analysis to analyse YouTube Italian videos concerning vaccination. *BMC Public Health* 20, 1 (2020), 1–9.

[10] Cees GM Snoek, Marcel Worring, and Arnold WM Smeulders. 2005. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th annual ACM international conference on Multimedia*. 399–402.

[11] Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. 2017. A survey of multimodal sentiment analysis. *Image and Vision Computing* 65 (2017), 3–14.

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

[13] Lei Zhang, Shuai Wang, and Bing Liu. 2018. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8, 4 (2018), e1253.