Multimodal Sentiment Analysis: Project Proposal

Lucía Prado - Andrés Martínez
Javier Ríos - Pablo García Molina

COMILLAS
UNIVERSIDAD PONTIFICIA

April 2024

1 Research Summary

In classic sentiment analysis systems, typically only one modality is considered to ascertain a user's sentiment toward a subject. However, as the landscape of online expression expands to encompass various media forms, including videos, there arises a need for more sophisticated approaches. Multimodal Sentiment Analysis (MSA) [2] emerges as a response to this shift, blending machine learning and deep learning advancements to accommodate multiple modalities such as text, visuals, and auditory cues.

Unlike conventional methods, MSA acknowledges the richness of expression across diverse media forms, recognizing that each modality offers unique insights into user sentiment. This approach, used for over a decade now (for example, in this study that analyses Movie Reviews using text, audio and video [5]), acknowledges, as well, the fact that the information from different sources may differ in its subjective meaning (f.e. sarcasm)

Visual features, for instance, play a pivotal role by succinctly conveying information that might be cumbersome to express through text alone. These features enrich sentiment prediction, allowing for a more accurate understanding of associated sentiments. Moreover, MSA explores various combinations of modalities, from speech+image to image+text, providing a holistic view of user sentiment.

Systems leveraging two modalities, termed bimodal sentiment analysis systems, and those incorporating all three, referred to as trimodal sentiment analysis systems, underscore the versatility of MSA in capturing nuanced expressions. The diverse sources of information, including audio, visual, and textual data, present both challenges and opportunities, driving MSA to enhance the accuracy and depth of sentiment understanding across different contexts and media types.

2 Key Ideas

2.1 Multimodal Feature Fusion

In order to process the information from the multiple sources, they must undergo a fusion process to be analyzed together [4]. A variety of techniques may be employed to carry out this process, categorized into three main approaches:

- Early Feature-Based Approaches
- Medium-Term Model-Based Fusion
- Model Based on Decision Fusion

2.2 Sentiment discrepancy and polarization

Two major problems, yet to solve, arise when trying to use a multi-model approach. For instance, when dealing with data from different sources they might indicate opposite sentiments due to phenomena such as sarcasm.

Added to this, sentiment-related features tend to not appear related enough just because the are originated from different sources. This two issues are mainly tackled by the methodology proposed "Hybrid cross-modal interaction learning for multi-modal sentiment analysis" [1] Briefly summarizing, the following techniques are exploited:

- Contrastive learning interactions
- Cross-modality predictions

2.3 Challenges due to Multiple Modalities

The incorporation of multiple data modalities (text, audio, visual) for comprehensive sentiment analysis entails a series of challenges to take into consideration:

- Lack of large, diverse datasets with high annotation accuracy.
- Difficulty in the detection of hidden emotions such as sarcasm, contextual and complex emotions.
- Challenges in analyzing video data due to noise, low-resolution, and complexity.
- Challenges in analyzing text data due to mixed languages and emotions.
- Continuous improvement of MSA techniques with the possibility of developing models with human-like sentiment analysis capabilities.

3 Project plan

3.1 Datasets

Both the main datasets and the models used in state of the art MSA have been obtained from Gandhi et al. [2], which compares them as well.

- IEMOCAP: audio, video, text, facial expression, and posture data, categorized into ten emotions. Limited actors and incomplete emotion categories.
- CMU-MOSEI: YouTube videos categorized into emotion and sentiment annotations in text, visual, and sound modalities. Limited emotion representation and subjective bias.
- Multi-ZOL: Dataset for bimodal sentiment classification of images and text, covering various mobile phone reviews. Noise in data and limited applicability.

3.2 Models

- MultiSentiNet-Att - incorporates LSTM for text information and VGG for image feature extraction using cross-modal attention mechanism to assign weights to sentiment-related word vectors. Employs multi-layer perceptron for sentiment analysis.
- HCIL - The model presented in "Hybrid cross-modal interaction learning for multimodal sentiment analysis" [1], sets new mechanics in feature-integration combining Attention, Transformers and LSTMs among others.
- MAG-BERT. The Multimodal Adaptation Gate for Bert (MAG-BERT) uses the adaptation gate mechanism at the BERT backbone to improve the RAVEN on aligned multimodal data.
- MUTA-Net. The modal-utterance-temporal attention network with multimodal sentiment loss (MUTA-Net) constructs global and local feature relationship structures to learn discriminative multi-relation representations.

3.3 Evaluation

We have decided to further explore the problem of MSA regarding features discrepancy. This problem mainly arise given the different natures of the data. We want to try and discover what techniques are currently being used to try tackle this issue, implement them and rank them. Finally, if time allows us and we are able to, we will try to propose some minor improvements.

3.4 Milestones

In order to complete the project, the work will be assessed by the accomplishment of two objectives in consecutive weeks:

1. April 12th: By the end of the first week, we will have decided upon the information datasets to use as well as the methods and models to implement.
2. April 19th: The models should be fully developed, and the evaluation method must be elected, in order to compare (in the days until the final deadline, if not before) the different models, as well as the conclusions

References

[1] Fu, Y., Zhang, Z., Yang, R., and Yao, C. Hybrid cross-modal interaction learning for multimodal sentiment analysis. *Neurocomputing 571* (2024), 127201.

[2] Gandhi, A., Adhvaryu, K., Poria, S., Cambria, E., and Hussain, A. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion 91* (2023), 424–444.

[3] Gupta, S. Sentiment analysis: Concept, analysis and applications. *Towards data science 7*, 06 (2018).

[4] Kaur, R., and Kautish, S. Multimodal sentiment analysis: A survey and comparison. *Research anthology on implementing sentiment analysis across multiple disciplines* (2022), 1846–1870.

[5] Wollmer, M., Weninger, F., Knaup, T., Schuller, B., Sun, C., Sagae, K., and Morency, L.-P. Youtube movie reviews: Sentiment analysis in an audio-visual context. *IEEE Intelligent Systems* (May 2013).