CSE474 Report

Name - Maisoon Tasnia
ID - 20301076
Sec - 01

Introduction

Multimodal sentiment analysis is a field of study that aims to analyze and interpret human sentiment by integrating information from multiple modalities, such as text, audio, video, and images. It goes beyond traditional text-based sentiment analysis by considering the rich and complementary cues present in various modalities.

The primary goal of multimodal sentiment analysis is to understand and interpret the emotions, attitudes, and opinions expressed by individuals in different media formats. By leveraging multiple modalities, it offers a more comprehensive and nuanced understanding of human sentiment.

The integration of multiple modalities enables the analysis of non-verbal cues like facial expressions, tone of voice, gestures, and visual context, which play a crucial role in conveying emotions and sentiments. For example, in a video, the facial expressions and gestures of a person can provide additional cues about their sentiment that cannot be captured by text alone.

Multimodal sentiment analysis has various applications in several domains. In the field of social media analysis, it can be used to extract sentiment from posts, comments, and multimedia content shared on platforms like Twitter, Facebook, or YouTube. In customer feedback analysis, it can help businesses understand customer sentiment expressed through various channels, including text reviews, audio recordings, and images.

To perform multimodal sentiment analysis, various machine learning and deep learning techniques are employed. These techniques involve the fusion of information from different modalities and the development of models capable of extracting sentiment-related features from each modality. The fusion can be early fusion (combining modalities at the input level), late fusion (combining modalities at the decision level), or hybrid fusion (combining modalities at multiple levels).

Challenges in multimodal sentiment analysis include handling data heterogeneity, aligning and synchronizing different modalities, handling missing or noisy data, and designing effective fusion strategies to leverage the complementary information from multiple modalities.

In summary, multimodal sentiment analysis is an interdisciplinary field that combines techniques from natural language processing, computer vision, audio processing, and machine learning to understand and interpret sentiment expressed across multiple modalities. By incorporating multimodal cues, it enhances the accuracy and richness of sentiment analysis, enabling

deeper insights into human emotions and opinions in various contexts.

Literature review on Multimodal sentiment analysis to explore the structure of emotions

Multimodal sentiment analysis is a field of research that seeks to explore the structure of emotions by analyzing data from multiple modalities, such as text, speech, and images. The goal is to understand the underlying patterns and relationships between the modalities and identify the key factors that influence the emotional content. Researchers have proposed various approaches for analyzing the structure of emotions, including deep learning-based models and multimodal feature extraction techniques. These approaches have shown significant improvements in sentiment analysis performance and have helped to uncover the complex interplay between different modalities in conveying emotional information. Understanding the structure of emotions can have important implications for a range of applications, from improving mental health to enhancing the accuracy of emotion detection in social media and other platforms.

Articles

1. A survey of multimodal sentiment analysis

Link`https://www.sciencedirect.com/science/article/abs/pii/S0262885617301191`

This paper provides an overview of multimodal sentiment analysis, including its definition, challenges, and applications. The paper reviews various approaches to combining different modalities, such as text, audio, and video, to improve sentiment analysis. The author also discusses the importance of addressing the affective dimensions of nonverbal communication, such as facial expressions, gestures, and body language. Furthermore, the paper provides insights into the applications of multimodal sentiment analysis in areas such as mental health, social media, and education. The author reports that multimodal sentiment analysis has achieved higher accuracy in

sentiment classification compared to single-modal approaches. However, the author notes that there are still several challenges in this field, including the lack of annotated multimodal datasets and the need for robust models that can handle different modalities. The paper concludes with a discussion of future research directions in multimodal sentiment analysis.

2. Multimodal Sentiment Analysis To Explore the Structure of Emotions

linkhttps://dl.acm.org/doi/abs/10.1145/3219819.3219853

This paper aims to explore the structure of emotions using multimodal sentiment analysis. The authors tackle the problem of identifying the underlying structure of emotions and how they are related to each other, using a multimodal approach that considers both textual and visual information. Their approach involves building a multimodal dataset that combines text, image, and audio features, and using dimensionality reduction techniques to extract latent emotional factors. The authors reported that their approach was able to capture the underlying structure of emotions, and identified six primary emotional factors: positive valence, negative valence, arousal, dominance, attentiveness, and formality. The results also showed that the multimodal approach outperformed unimodal approaches that only considered textual or visual information. Overall, the paper presents a promising approach to exploring the structure of emotions using multimodal sentiment analysis, with potential applications in fields such as psychology, marketing, and human-computer interaction.

3. A Co-Memory Network for Multimodal Sentiment Analysis

linkhttps://dl.acm.org/doi/abs/10.1145/3209978.3210093

This paper tackles the problem of multimodal sentiment analysis using

a co-memory network approach that integrates multiple modalities of input data. The authors propose a deep neural network architecture that can jointly model textual and visual features, using a co-memory module that allows information to be exchanged between the modalities. Their approach also incorporates a co-attention mechanism that focuses on the most important features of each modality. The authors evaluated their approach on several benchmark datasets, showing that their co-memory network outperformed state-of-the-art methods that only considered a single modality. Additionally, they conducted an ablation study to show the effectiveness of the co-memory module and the co-attention mechanism. Overall, the paper presents a promising approach to multimodal sentiment analysis that can integrate multiple sources of information to improve performance.

4. Cross-modality Consistent Regression for Joint Visual-Textual Sentiment Analysis of Social Multimedia

link`https://dl.acm.org/doi/abs/10.1145/2835776.2835779`

This paper addresses the problem of joint visual-textual sentiment analysis of social multimedia by proposing a novel cross-modality consistent regression (CMCR) model. The authors' approach involves using a deep neural network to extract features from both visual and textual modalities, followed by a multi-task learning framework to jointly predict sentiment scores. The key contribution of their work is the introduction of a consistency constraint, which aims to align the predictions of each modality. The authors evaluated their model on two publicly available datasets and showed that their approach outperformed several state-of-the-art methods. Furthermore, they conducted an ablation study to demonstrate the effectiveness of the consistency constraint. Overall, the paper presents a promising approach to joint visual-textual sentiment analysis that can effectively leverage information from multiple modalities.

5. MultiSentiNet: A Deep Semantic Network for Multimodal Sentiment Analysis

linlhttps://dl.acm.org/doi/abs/10.1145/3132847.3133142

This paper addresses the problem of multimodal sentiment analysis by proposing a novel deep neural network model called MultiSentiNet. The authors' approach involves using multiple convolutional neural networks (CNNs) to extract features from different modalities, followed by a semantic network that integrates these features and generates a joint sentiment representation. The key contribution of their work is the use of a semantic network, which allows the model to capture more nuanced relationships between the input modalities. The authors evaluated their model on two publicly available datasets and showed that their approach outperformed several state-of-the-art methods. Furthermore, they conducted an ablation study to demonstrate the effectiveness of the semantic network. Overall, the paper presents a promising approach to multimodal sentiment analysis that can effectively leverage information from multiple modalities and capture more nuanced relationships between them.

6. Benchmarking Multimodal Sentiment Analysis

linkhttps://link.springer.com/chapter/10.1007/978-3-319-77116-8_13

This paper addresses the problem of benchmarking multimodal sentiment analysis systems. The authors argue that the lack of standardized evaluation protocols makes it difficult to compare different multimodal sentiment analysis methods, hindering the progress of the field. Therefore, they propose a comprehensive benchmarking framework that includes a dataset with multimodal inputs and a set of evaluation metrics. The authors evaluated several existing multimodal sentiment analysis methods using their proposed framework and provided detailed performance analyses. Additionally, they proposed a new method based on deep neural networks that achieved state-

of-the-art performance on their dataset. The key contribution of their work is the proposed benchmarking framework, which enables fair comparison and evaluation of multimodal sentiment analysis methods. Overall, the paper provides an important contribution to the field of multimodal sentiment analysis and provides a useful resource for researchers and practitioners.

7. Towards multimodal sentiment analysis: harvesting opinions from the web

linkhttps://dl.acm.org/doi/abs/10.1145/2070481.2070509

This paper addressed the challenge of analyzing sentiment using multiple modalities, such as text, audio, and video. The authors proposed a method that combines textual and non-textual cues to infer sentiment in social media data. The approach involved first extracting features from the different modalities, then applying a sentiment classifier to each modality separately. Finally, the outputs of the classifiers were combined using a fusion method to obtain a final sentiment score. The authors evaluated their approach on two publicly available datasets and showed that their multimodal approach outperformed traditional approaches that only used a single modality. They also demonstrated the usefulness of incorporating non-textual modalities by showing that the performance of their approach increased when audio and video data were included. This paper provides a promising avenue for future research in multimodal sentiment analysis.

8. Image-Text Multimodal Emotion Classification via Multi-View Attentional Network

linkhttps://ieeexplore.ieee.org/abstract/document/9246699

This paper tackled the challenge of classifying emotions from both images and text data. The authors proposed a novel approach that utilizes a multi-view attentional network to capture the interactions between the two

modalities. The approach involves first extracting visual and textual features from images and their corresponding captions, respectively. Then, a multi-view attentional network is used to dynamically learn the most informative features from both modalities. Finally, a classifier is applied to predict the emotion class. The authors evaluated their approach on two publicly available datasets and showed that their approach outperformed state-of-the-art approaches that only used a single modality. They also demonstrated the usefulness of the multi-view attentional network by showing that it can effectively capture the interactions between the two modalities. This paper provides a promising approach for future research in multimodal emotion classification.

Repositories

Multimodal Sentiment Analysis

Link`https://github.com/lemei/unimse`
UniMSE is a Python code repository developed by Lei et al. (2019) that provides a framework for sentiment analysis based on unimodal, multiscale embeddings. The authors utilized several resources, including word embeddings, sentiment knowledge graphs, and sentiment dictionaries, to develop their approach. The framework includes code for data preprocessing, creating unimodal multiscale sentiment embeddings, and training and evaluating sentiment classifiers. The authors evaluated their approach on several benchmark datasets and reported improved performance compared to baseline methods. The UniMSE framework and code repository are valuable resources for industry and academic research, providing a new approach to sentiment analysis that can be customized to fit specific needs. Additionally, the use of sentiment knowledge graphs and dictionaries can enhance the interpretability of sentiment analysis results, which is critical for applications that require explainable AI.

Methodology

Data Collection: Gather multimodal data that includes text, audio, video, or image sources. This data can be collected from social media platforms, online reviews, customer feedback, or other relevant sources.

Data Preprocessing: Clean and preprocess the collected data to ensure consistency and remove noise. This step may involve text normalization, noise removal from audio, video preprocessing, and image processing techniques such as resizing or cropping.

Modality-specific Feature Extraction: Extract features from each modality to represent the sentiment-related information. For text, this may involve techniques like word embeddings or bag-of-words representations. Audio features can include pitch, intensity, or spectral features. Video features may include facial expressions, gestures, or body movements. Image features can involve visual cues such as color, texture, or shape.

Modality Fusion: Combine the features extracted from different modalities to create a unified representation. Fusion techniques can be applied at various levels, such as early fusion (combining features at the input level), late fusion (combining predictions at the decision level), or hybrid fusion (combining features at multiple levels).

Sentiment Analysis Model: Develop a sentiment analysis model that takes the multimodal representation as input and predicts sentiment labels. This model can be based on various machine learning or deep learning techniques, such as support vector machines (SVM), recurrent neural networks (RNN), convolutional neural networks (CNN), or transformers.

Training and Evaluation: Split the multimodal data into training and testing sets. Train the sentiment analysis model on the training data and evaluate its performance on the testing data. Evaluation metrics such as accuracy, precision, recall, and F1-score can be used to assess the model's performance.

Fine-tuning and Optimization: Fine-tune the model by adjusting hyperparameters and optimizing the model architecture to improve its performance. This step may involve techniques like cross-validation, grid search, or Bayesian optimization.

Deployment and Application: Once the sentiment analysis model achieves satisfactory performance, it can be deployed for real-world applications. This may involve integrating the model into a larger system or using it for sentiment analysis tasks on new data.

Discussion

Enhanced Sentiment Understanding: By incorporating multiple modalities, multimodal sentiment analysis allows for a more comprehensive understanding of sentiment. Each modality contributes unique information that can help capture subtle nuances and context-specific cues related to sentiment. For example, facial expressions and body language in video data, tone of voice in audio data, and specific words or phrases in text data can all provide valuable insights into sentiment.

Complementary Information: Different modalities often complement each other in capturing sentiment. For instance, textual data provides explicit opinions and expressions, while visual and auditory cues provide implicit emotional cues that may not be evident in text alone. By combining these modalities, multimodal sentiment analysis can provide a more holistic and accurate representation of sentiment.

Modality Alignment: One of the challenges in multimodal sentiment analysis is aligning different modalities and their associated data. Ensuring temporal, spatial, or semantic alignment between modalities is crucial for accurate interpretation and fusion of information. Techniques such as synchronization algorithms, time-stamping, and alignment models are employed to address this challenge.

Fusion Strategies: Multimodal sentiment analysis involves the fusion of information from different modalities. Fusion strategies can be categorized as early, late, or hybrid fusion, depending on when and how the fusion occurs in the analysis pipeline. Early fusion combines modalities at the feature level, whereas late fusion combines predictions or decisions made by individual modality-specific models. Hybrid fusion combines features at multiple levels. The choice of fusion strategy depends on the specific task and the characteristics of the data.

Conclusion

In conclusion, multimodal sentiment analysis is a rapidly evolving field

that aims to enhance sentiment understanding by incorporating multiple modalities, such as text, audio, video, and images. By leveraging the complementary information provided by these modalities, multimodal sentiment analysis offers a more comprehensive and accurate representation of sentiment. However, multimodal sentiment analysis also presents challenges. Handling data heterogeneity, ensuring modality alignment, and developing effective fusion techniques are areas that require further research and development. Ethical considerations, such as privacy protection and bias mitigation, should also be addressed to ensure responsible and fair use of multimodal data.

Despite these challenges, multimodal sentiment analysis holds great potential for advancing sentiment analysis tasks and providing valuable insights into human emotions and opinions. Continued research and innovation in this field will contribute to improved sentiment analysis techniques and applications in various domains, ultimately enhancing our understanding of human sentiment