Title: Multimodal sentiment analysis

URL: https://en.wikipedia.org/wiki/Multimodal_sentiment_analysis

PageID: 57687371

Categories: Category:Affective computing, Category:Machine learning, Category:Multimodal interaction, Category:Natural language processing, Category:Social media

Source: Wikipedia (CC BY-SA 4.0).

Multimodal sentiment analysis is a technology for traditional text-based sentiment analysis , which includes modalities such as audio and visual data. [1] It can be bimodal, which includes different combinations of two modalities, or trimodal, which incorporates three modalities. [2] With the extensive amount of social media data available online in different forms such as videos and images, the conventional text-based sentiment analysis has evolved into more complex models of multimodal sentiment analysis, [3] [4] which can be applied in the development of virtual assistants , [5] analysis of YouTube movie reviews, [6] analysis of news videos, [7] and emotion recognition (sometimes known as emotion detection) such as depression monitoring, [8] among others.

Similar to the traditional sentiment analysis , one of the most basic task in multimodal sentiment analysis is sentiment classification, which classifies different sentiments into categories such as positive, negative, or neutral. [9] The complexity of analyzing text, audio, and visual features to perform such a task requires the application of different fusion techniques, such as feature-level, decision-level, and hybrid fusion. [3] The performance of these fusion techniques and the classification algorithms applied, are influenced by the type of textual, audio, and visual features employed in the analysis. [10]

Features

Feature engineering , which involves the selection of features that are fed into machine learning algorithms, plays a key role in the sentiment classification performance. [10] In multimodal sentiment analysis, a combination of different textual, audio, and visual features are employed. [3]

Textual features

Similar to the conventional text-based sentiment analysis , some of the most commonly used textual features in multimodal sentiment analysis are unigrams and n-grams , which are basically a sequence of words in a given textual document. [11] These features are applied using bag-of-words or bag-of-concepts feature representations, in which words or concepts are represented as vectors in a suitable space. [12] [13]

Audio features

Sentiment and emotion characteristics are prominent in different phonetic and prosodic properties contained in audio features. [14] Some of the most important audio features employed in multimodal sentiment analysis are mel-frequency cepstrum (MFCC) , spectral centroid , spectral flux , beat histogram, beat sum, strongest beat, pause duration, and pitch . [3] OpenSMILE [15] and Praat are popular open-source toolkits for extracting such audio features. [16]

Visual features

One of the main advantages of analyzing videos with respect to texts alone, is the presence of rich sentiment cues in visual data. [17] Visual features include facial expressions , which are of paramount importance in capturing sentiments and emotions , as they are a main channel of forming a person's present state of mind. [3] Specifically, smile , is considered to be one of the most predictive visual cues in multimodal sentiment analysis. [12] OpenFace is an open-source facial analysis toolkit available for extracting and understanding such visual features. [18]

Fusion techniques

Unlike the traditional text-based sentiment analysis , multimodal sentiment analysis undergo a fusion process in which data from different modalities (text, audio, or visual) are fused and analyzed together. [3] The existing approaches in multimodal sentiment analysis data fusion can be grouped into three main categories: feature-level, decision-level, and hybrid fusion, and the performance of the sentiment classification depends on which type of fusion technique is employed. [3]

Feature-level fusion

Feature-level fusion (sometimes known as early fusion) gathers all the features from each modality (text, audio, or visual) and joins them together into a single feature vector, which is eventually fed into a classification algorithm. [19] One of the difficulties in implementing this technique is the integration of the heterogeneous features. [3]

Decision-level fusion

Decision-level fusion (sometimes known as late fusion), feeds data from each modality (text, audio, or visual) independently into its own classification algorithm, and obtains the final sentiment classification results by fusing each result into a single decision vector. [19] One of the advantages of this fusion technique is that it eliminates the need to fuse heterogeneous data, and each modality can utilize its most appropriate classification algorithm . [3]

Hybrid fusion

Hybrid fusion is a combination of feature-level and decision-level fusion techniques, which exploits complementary information from both methods during the classification process. [6] It usually involves a two-step procedure wherein feature-level fusion is initially performed between two modalities, and decision-level fusion is then applied as a second step, to fuse the initial results from the feature-level fusion, with the remaining modality . [20] [21]

Applications

Similar to text-based sentiment analysis, multimodal sentiment analysis can be applied in the development of different forms of recommender systems such as in the analysis of user-generated videos of movie reviews [6] and general product reviews, [22] to predict the sentiments of customers, and subsequently create product or service recommendations. [23] Multimodal sentiment analysis also plays an important role in the advancement of virtual assistants through the application of natural language processing (NLP) and machine learning techniques. [5] In the healthcare domain, multimodal sentiment analysis can be utilized to detect certain medical conditions such as stress , anxiety , or depression . [8] Multimodal sentiment analysis can also be applied in understanding the sentiments contained in video news programs, which is considered as a complicated and challenging domain, as sentiments expressed by reporters tend to be less obvious or neutral. [24]

References