Meeting Note

Name: Henry Lin
Date: 2024/06/24 – 2024/06/28

Summary
Emotion-based recommendation system [1, 2, 3, 4, 6].
Multi-modal emotion detection using Transformer [5].
Multi-modal emotion recognition [7].

Problems
Most of the traditional recommendation system recommend contents based on the users' historical behaviors, but doesn't consider the users' emotion [1].

How do we use recommendation system to control the metal status of users [2] or build a highly personalized recommendation system [3, 4]?

What are the SOTA multi-modal emotion recognition techniques [7]?

Details

Priya et al. [1] introduced an emotion-based recommendation system that can learn the emotions of the user from the camera and recommend music, videos, and other multi-media contents to users. They first use the camera to read the user's image, and use segmentation functions to obtain the facial images. Then, they utilize the convolutional neural network (CNN) to extract the feature vectors. To obtain the real time performance and reduce the time complexity, in the segmentation stage, they only remain the eyes and mouse area. After all, the system will play the music based on the predicted emotion.

Thenmozhi et al. [2] build an emotion-based recommendation system that aims to help users light up their mood for good. They use the OpenCV HAAR-Cascade to recognize the face and evaluate expressions and mood. They divided the metal status into the following three different kinds: happy, sad, and anger, then give different contents suggestion to users. Once the contents are suggested to the users, they just have to use their voice to select one item from the recommendation list.

In this study, the recommendation system is just a simple conditional based system, i.e., if the predicted result is "sad", the system will recommendation

some pre-stored funny contents to the user. And the action of the “speech recognition” part is used to select the item from the recommendation list, not used to recognize the user’s emotion.

Babanne et al. [3] modify a framework of video contents recommendation system to make better recommendations based on the current emotional information of user. Their approach is like the first one, to achieve the real-time performance, they extract the facial features through video inputs from users. In addition, they combine the Local Binary Patterns Histogram (LBPH) with HAAR Cascade to efficiently extract the features. On the other hand, they not just detect the emotional status from the facial image, but also extract age and gender information from it. This provides the recommendation system more information to make better suggestions.

Costa et al. [4] proposed a new component in recommendation system, which considers the user’s mental status, performance and social networking trends. At the part of recommendation system, they introduce a textual feature extractor to extract the most relative information from the daily news and filter the irrelative items from the list, then the remains will be stored into a database. Once the agent decides what news to recommend, it will choose the relative items from the database.

In this paper, the authors didn’t clearly explain how their architecture works (e.g., what does it mean to “consider the social trends”), but just gave a rough overview.

Ju et al. [5] introduce a Transformer-based model to detect the user’s mental status (emotions) and generate the label set. This model takes text, a sequence of images and audio as the input, and output a set of labels which represents the user’s emotions. First, they use a cross-modal encoder capture the cross-modal interaction between different modalities. The data of different modalities are linearly transformed into sequences of same dimensions and inputted into the Transformer encoder. Then the decoder outputs the label set (e.g., “happy, surprise”). As the authors mention, the conventional multi-modal multi-label emotion detection tasks usually perform the binary classification of each emotion category, the contribution of this paper is that they use a generative model to perform the multi-label classification task.

This gives me new ideas to design the Transformer-Based Multi-Modal Classifier (TBMMC), such as replace the feature extractors with linear layers.

Kim et al. [6] utilized the genetic algorithm to classify the emotion status of speech data, they divide the human emotion status into six classes (e.g. neutral, happy, sad, angry, surprised, and bored) and obtain the accuracy of 86.98%. At the part of media contents, image and music, in this study, they use factor analysis, correspondence analysis, and Euclidean distance to classify their emotion classes. The proposed system can use music information and each user's emotion history to recommend music that is appropriate for the user's current mood according to their speech emotion information.

Lian et al. [7] did a survey on the multi-modal emotion recognition techniques of these days. They first introduced the popular datasets in this field, such as IEMOCAP, MOUD, ICT-MMMO, etc. Then, the feature extraction techniques of three modalities (e.g., speech, text, and facial image) are introduced. For the speech feature, there are two major kinds of features, one is hand-crafted feature and the other is deep feature, the former take speech data as signals that are designed by people based on prior knowledge and professional experience, such as prosodic features, voice quality features, and spectral features; the latter is the features that extracted by the deep learning techniques (e.g., wav2vec, WavLM). For the textual features, the authors mentioned some traditional natural language processing (NLP) methods, such as bag-of-words (BoW) and pretrained models like BERT, that are used to classify the emotion of textual data. For the facial feature, some conventional methods such as Scale Invariant Feature Transform (SIFT), Active Shape Model (ASM), and deep learning methods such as 3DCNN, C-LSTM, T-LSTM are mentioned. After the features are extracted, the next step is features fusion. In the section 4, six fusion methods are introduced.

In the introduction section, we found that there is limited research on textual emotion recognition by comparison, most research focus on the facial and speech emotion recognition.

Conclusions

All the above proposed recommendation systems [1, 2, 3, 4, 6] didn't completely consider the information which generated by users (e.g., text typed by user, or even the EEG signal) to recognize the user's mental status, most of them are just use CNNs to extract features of speech or facial image, classify the mental status and make recommendation suggestions. This gives us a chance to use the TBMMC to build a general architecture of recommendation system based on multi-modal emotion classification.

As far as we know, there is still limited research on how many information a mankind has (e.g., a human has information of EEG signal, acoustic fingerprint, emotion, personality, color, hair color, salary, shopping history, faith, etc.). We can first do research to study what and how much information can we collect from users, and then we can use the information to identify an individual, recommend multi-media contents, and even build a crime prediction system to predict how likely is it for a human to commit a crime. Furthermore, we can utilize this information to control the human's behaviors through the recommendation system.

References

[1] P.Priya dharshini, S.Sowmya, J. Gayathri. Emotion Based Recommendation System for Various Applications. IJRASET, 2019.

[2] Thenmozhi T, Geerisha Jain. Mood-Up: Emotion Based Recommendation System with Face and Speech Recognition. IJCSMA, 2022.

[3] Vanita Babanne, Mrunal Borgaonkar, Mrunali Katta, Prajakta Kudale, Vaishnavi Deshpande. Emotion based Personalized Recommendation System. IRJET, 2020.

[4] Costa, Hernani & Macedo, Luis. (2013). Emotion-Based Recommender System for Overcoming the Problem of Information Overload. Communications in Computer and Information Science. 365. 10.1007/978-3-642-38061-7_18.

[5] Xincheng Ju, Dong Zhang, Junhui Li, Guodong Zhou; (2020). Transformer-based Label Set Generation for Multi-modal Multi-label Emotion Detection. Proceedings of the 28th ACM International Conference on Multimedia. doi:10.1145/3394171.3413577

[6] Kim TY, Ko H, Kim SH, Kim HD. Modeling of Recommendation System Based on Emotional Information and Collaborative Filtering. Sensors (Basel). 2021 Mar 12;21(6):1997. doi: 10.3390/s21061997. PMID: 33808989; PMCID: PMC7999638.

[7] Lian, H.; Lu, C.; Li, S.; Zhao, Y.; Tang, C.; Zong, Y. A Survey of Deep Learning-Based Multimodal Emotion Recognition: Speech, Text, and Face. Entropy 2023, 25, 1440. https://doi.org/10.3390/e25101440