Mário Francisco
Costa Silva

Avaliação de Áudio de Dados Afetivos em Sessões de Videoconferência

Audio Evaluation of Affective Data in Videoconference Sessions

Mário Francisco
Costa Silva

Avaliação de Áudio de Dados Afetivos em Sessões de Videoconferência

Audio Evaluation of Affective Data in Videoconference Sessions

"*Just like we can understand speech and machines can communicate in speech, we also understand and communicate with humor and other kinds of emotions. And machines that can speak the language of emotions are going to have better, more effective interactions with us*"

— MIT Sloan professor Erik Brynjolfsson

Universidade de Aveiro
2023

Mário Francisco
Costa Silva

Avaliação de Áudio de Dados Afetivos em Sessões de Videoconferência

Audio Evaluation of Affective Data in Videoconference Sessions

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Engenharia Informática, realizada sob a orientação científica da Doutora Susana Manuela Martinho dos Santos Baía Brás, Professora Investigadora no Instituto de Engenharia Eletrónica e Telemática de Aveiro do Departamento de Eletrónica, Telecomunicações e Informática da Universidade de Aveiro, e do Doutor Ilídio Castro Oliveira, Professor Auxiliar do Departamento de Eletrónica, Telecomunicações e Informática
.

Palavras Chave

Computação Afetiva, Processamento de Voz, Reconhecimento de Emoção da Fala, Aprendizagem Automática

Resumo

Esta disseratação apresenta um estudo abrangente do Reconhecimento de Emoção de Fala (REF) usando abordagens tradicionais de aprendizagem automática e aprendizagem profunda. A principal contribuição deste trabalho é o desenvolvimento e avaliação de dois modelos em múltiplos datasets. Além de explorar o impacto de diferentes conjuntos de características do áudio e metodologias de validação, também investigamos a importância das técnicas de pré-processamento de áudio e seu efeito no desempenho do modelo. Por meio de estudos experimentais, desenvolvemos dois modelos para REF: um modelo eXtreme Gradient Boosting (XGBoost) para a abordagem tradicional utilizando um vetor unidimensional de 33 características do áudio e um modelo ResNet50 ajustado usando imagens de espectrograma para aprendizagem profunda. Estes modelos alcançaram precisões de 60,69% e 58,24%, respectivamente, para validação cruzada estratificada de 5 vezes no dataset Interactive Emotional Dyadic Motion Capture (IEMOCAP). Além disso, o modelo de aprendizagem profunda superou o modelo tradicional na avaliação cruzada de datasets devido à sua maior capacidade de extração e generalização de recursos, enquanto o modelo tradicional é mais adequado para aplicações em tempo real devido à sua velocidade de processamento mais rápida. Além disso, uma análise detalhada dos dados usando várias estratégias de estratificação levou à identificação de um conjunto de condições para o IEMOCAP que melhorou o desempenho geral dos modelos na avaliação cruzada de datasets. Também foi desenvolvido um pipeline que automatiza o processo de REF para classificação em tempo real ou offline, criando segmentos de áudio com uma determinada duração e classificando as emoções presentes neles usando os modelos desenvolvidos. No geral, esta disseratação fornece uma base sólida para pesquisas futuras no desenvolvimento de modelos REF mais robustos e precisos, oferecendo uma implementação abrangente e processo de pensamento, juntamente com conclusões e interpretações dos resultados obtidos. As nossas conclusões contribuem para o crescente corpo de pesquisa de REF e fornecem informações valiosas para investigadores e profissionais da área.

Keywords Affective Computing, Voice Processing, Speech Emotion Recognition, Machine Learning

Abstract

This dissertation presents a comprehensive study of Speech Emotion Recognition (SER) using traditional machine learning and deep learning approaches. The main contribution of this work is the development and evaluation of two models on multiple datasets. In addition to exploring the impact of different feature sets and validation methodologies, we also investigate the importance of audio preprocessing techniques and their effect on model performance. Through experimental studies, we developed two models for SER: an eXtreme Gradient Boosting (XGBoost) model for the traditional approach utilizing a 1-dimensional vector of 33 audio features, and a fine-tuned ResNet50 model using spectrogram images for deep learning. These models achieved accuracies of 60.69% and 58.24%, respectively, for 5-fold stratified cross-validation on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset. Moreover, the deep learning model outperformed the traditional model in the cross-dataset evaluation due to its higher feature extraction and generalization capabilities, while the traditional model is more suitable for real-time applications due to its faster processing speed. Furthermore, a detailed analysis of the data using several stratification strategies led to the identification of a set of conditions for IEMOCAP that improved the general performance of the models in cross-dataset evaluation. A pipeline was also developed that automates the SER process for both real-time and offline classification by creating voiced audio segments with a certain duration and classifying the emotions present in them using the developed models. Overall, this dissertation provides a solid foundation for future research in developing more robust and accurate SER models, by offering a comprehensive implementation and thought process, along with conclusions and interpretations of the obtained results. These findings contribute to the growing body of research on SER and provide valuable insights for researchers and practitioners in the field.