Literature Review and Summaries

Lecture on Emotion Analysis at University of Stuttgart
WS 2020/2021
https://www.emotionanalysis.de/
Roman Klinger

Authors: Thomas Bott; Adnan Ahmad; Ali Salaheddine; Isabelle Mohr; Sarina Meyer; Nishan Chatterjee; Alexander Christodoulou; Xanat Herrera; Eileen Wemmer; Iveta Senkane; Carlotta Quensel; Felix Bühler; Maximilian Wegge; Hanna Surjadi; Hasan Cengiz Oöztürk; Ema-Maria Zaviacicova; Leonie Lanfrit; Rajesh Baidya; Van Hoang; Nina Dörr; Pavan Mandava; Victoria Punstel; Alina Braitmaier; Jasvinder Singh; Meghdut Sengupta; Michael Göggelmann; Pavlos Musenidis; Milena Voskanyan; Wei Zhou; Faizan E Mustafa; Yannic Becker; Marius Maile; Christina Hitzl; Shawon Ashraf; Ching-Yi Chen;

Contents

Social Media

Thomas Bott

> Dua'a Al-Hajjar1, Afraz Z. Syed, (2015): Applying Sentiment and Emotion Analysis on Brand Tweets for Digital Marketing, IEEE AEECT. https://ieeexplore.ieee.org/abstract/document/7360592

Motivation

A lot of people express their thoughts and feelings about brands and companies on social media. Because of that, there is a growing interest in judging user opinions and emotional states on social media platforms for digital marketing. Thereby the main goal is to better understand mass opinions and perspectives about the brand and its launched products. Furthermore one can obtain a brand image which is useful for marketers to observe what customers specifically feel negative or positive about. The main research question of the paper is to investigate whether the accuracy of the recognition of sentiments and emotions for the benefit of digital marketing can be improved by following a combined approach of sentiment and emotion analysis.

Data

The authors of the paper collect 10000 English tweets from a 10 day period via Twitter4J for 10 technical brands: Apple, Google, Microsoft, Samsung, GE, IBM, Intel, Facebook, Oracle and HP. By doing that, they acquire 10 brand corpora, each containing 1000 tweets.

Method

For a brand corpus, they perform both sentiment and emotion analysis on each tweet by using dictionaries. For sentiment analysis they map adjectives and adverbs from a tweet to the SentiWordNet Lexicon and calculate a sentiment ratio value with a polarity. For emotion analysis they map adjectives, adverbs and hashtags to the NRC Hashtag Emotion Lexicon which contains Plutchik's 8 emotions. From the output of the lexicon, they take the maximum score with its associated emotion.
To combine the results from both analyses, they normalize the scores and divide Plutchik's emotions into positive and negative emotions. If both the polarity of the sentiment and the labeled emotion are positive (or negative), the prediction is seen as accurate. However, if the sentiment is positive and the emotion negative (or vice versa), the normalized scores are compared and the higher value dominates the prediction.
Like this, they assign a (polarity, emotion) label pair to each tweet. They evaluate the predictions from all three approaches (sentiment, emotion, combined) on a randomly extracted sample set containing 100 tweets and report average and detailed accuracy values.

Main Result

The main finding of the paper is that the approach of combining the predictions of sentiment and emotion analysis leads to improved results compared to the individual approaches (0.526 compared to 0.373 and 0.392). They emphasize that their approach might be specifically useful as an application in digital marketing where it is possible to obtain an improved image of a brand.

Critical Reflection, Limitations

They do not report how they annotate the tweets in the evaluation set. It would be interesting to see if domain specific dictionaries lead to improved results. A next step could be to take the context of a tweet into account to further improve results and get a more detailed brand image.

Adnan Ahmad

> Bollen Johan, Huina Mao, and Alberto Pepe. ... (2011): "Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena.", Proceedings/Proceedings of the International AAAI Conference on Web and Social Media. Vol. 5. No. 1. https://ojs.aaai.org/index.php/ICWSM/article/view/14171

Motivation

The goal was to find that events in the social, political, cultural and economic sphere do have a significant, immediate and highly specific effect on the various dimensions of public mood in social media.

Data

- Public tweets broadcasted by Twitter users between August 1 and December 20, 2008, a total of 1.1M tweets containing mostly expressions of individual mood states.
- Fluctuations recorded by stock market and crude oil price indices and major events in media and popular culture such as the U.S. Presidential Election and Thanksgiving Day.

Method

- They perform a sentiment analysis of all public tweets.
- For every day in the timeline, they extract six dimensions of mood (tension, depression, anger, vigor, fatigue, confusion) using an extended version of the Profile of Mood States (POMS), a well-established psychometric instrument.
- They compare our results to fluctuations recorded by stock market and crude oil price indices and major events in media and popular culture, such as the U.S. Presidential Election of November 4, 2008 and Thanksgiving Day.
- They find that events in the social, political, cultural and economic sphere do have a significant, immediate and highly specific effect on the various dimensions of public mood.

Main Result

- They observe a spike in Depression and Confusion, and remarkably a sharp drop in Fatigue that started two days before election day. After the election, mood levels drop to nominal levels, except a significant spike in Vigour and a large drop in Fatigue. (Figure 1a)
- Second case on Thankgiving shows mood dimensions remain nearly at baseline levels with the exception of Vigour which spikes significantly on Thanksgiving Day indicating happy mood. (Figure 1b)
- They also determine whether the differences in mood levels are in fact statistically significant. They perform a Mann-Witney U-test over the time series observed within a period for each mood dimension.they showed that the difference in mood level is statistically significant.

Critical Reflection, Limitations

The time window is not large enough for such study. Also, multiple years should be included into the study to observe repetitive patterns for specific social events.

Ali Salaheddine

Jiabin Pan, Naixia Mou, Wenbao Liu (2019): Emotion Analysis of Tourists Based on Domain Ontology, ICDMML 2019: Proceedings of the 2019 International Conference on Data Mining and Machine. https://dl.acm.org/doi/abs/10.1145/3335656.3335701

Motivation

More and more people express their travel experience online by publishing comments, writing travel notes, uploading photos, etc. This offers large amounts of data that can be utilized for online commentary emotion analysis. So the main research goal of this work was to derive emotional information of tourists from tourism big data.

Data

The data used are comments on the so called Palace Museum which were aggregated from Ctrip (http://www.ctrip.com) China's biggest online travel agency. They used all comments which were posted in 2018 which accounts to approximately 15000 data items after data clean up. They obtained this data by writing their own web crawler. They did not only use the content of the comments for analysis but also used the date at which the comment was posted in order to detect monthly changes in the emotion.

Method

In order to retrieve the emotional information of the comments they created and implemented an emotion classifier and applied it to the aggregated data. First, they created a tourism domain ontology. Secondly, they used an NLP API to clean up the content of the comments. Then applied the comments on the ontology to extract the key features from the comment capturing its emotions. Thirdly, they created a dictionary for assessing the emotions of Chinese words. The extracted features are then applied to the dictionary and used to calculate the emotion score for each emotion.

Main Result

The emotion analysis of the palace museum delivered the following findings. Tourists find aspects like the scenery , amount of tourists visiting, and convenience more important and are less concerned about eating and entertainment and services and fees. Another interesting finding which was derived is that the worst travel experiences were in the months May, July, August and October due to the high number of holiday tourists.

Critical Reflection, Limitations

As the results show emotion analysis of tourism data is a very interesting idea and can be very beneficial for cities reliant on tourism in order to detect trends which tourists enjoy and dislike and allow the city to adapt accordingly relatively fast. One limitation I found is in the selection of the data. Only using comments from the travel agencies website can be biased as people are more inclined to comment on something when they had a negative experience. Therefore, analyzing this effect and more variety in the data collected could be possible improvements and next steps. Another limitation is that the emotion extraction and dealing with ambiguous words were very simplified. This could be due to not using an existing emotion model and solved by using an existing emotion model with proven methods.

Music

Isabelle Mohr

> Juan Sebastián Gómez Cañón, Estefanía Cano, Perfecto Herrera, Emilia Gómez, (2020): Transfer learning from speech to music: towards language-sensitive emotion recognition models, EUSIPCO. `https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9287548`

Motivation

There seem to be similarities between perception of emotions in music and in speech. These include shared taxonomies of emotion, shared processes in the brain and perhaps similar acoustic cues in both domains. Research in music psychology points out that emotion perception may be influenced by factors such as language and culture [1]. The authors thus attempt to discover whether inductive transfer learning can be used to create language-sensitive music emotion recognition (MER) models.

Data

The speech data used in this paper was Librispeech, an English speech recognition dataset, and AISHELL, a Mandarin speech dataset collected from different regions in China. As for labeled music data, the authors used the 4Q-emotion dataset, consisting of mainly Western popularly consumed music, and the CH-818 dataset consisting of Chinese pop music.

Method

Mel-spectograms were extracted as features from all datasets using the librosa package for Python. The authors designed a sparse convolutional autoencoder (SCAE) with rectified linear unit (ReLU) activations, which was used for pretraining (either in English or in Mandarin). Transfer learning was then implemented by replacing the decoder in the SCAE with three fully connected (FC) layers, thereby flattening the model, and adding three blocks of two FC layers, each block representing a classifier. The first classifier predicts quadrants (4 classes, one per quadrant), the second predicts arousal and the third predicts valence. The four quadrants make up emotions as follows: Q1: positive arousal & valence (happiness), Q2: positive arousal & negative valence (anger), Q3: negative arousal & valence (sadness), and Q4: negative arousal & positive valence (tenderness). The authors compare their model to a baseline CNN model (see paper for details).

Main Result

Features learnt during pretraining on speech are generally transferrable to music. In the baseline CNN, difficulties in both intra- and inter-linguistic settings show up as confusions between Q1 and Q2 (both with positive arousal) as well as Q3 and Q4 (both with negative arousal). Thus, arousal is more easily predicted (deduced from tempo and loudness) but valence not. SCAE reduces this confusion by improving quadrant, valence and arousal prediction, showing the benefits of pretraining with speech data and fine-tuning with music data. The intra-linguistic models (eng2eng/man2man) also outperform cross-linguistic models, showing preference for language-sensitive models.

Critical Reflection, Limitations

It may be problematic to conflate language and culture - should these be separately accounted for in future models? It would also be interesting to see models based on Eckman's basic emotions - the quadrant to emotion mapping seems to be lacking in complexity.

[1] Note: This paper also included a survey and analysis concerning native language which is not mentioned in this review (because of space constraints) but is nonetheless very interesting. Please refer to the paper if interested.

Sarina Meyer

> Daniela F. Milon-Flores, Jose Ochoa-Luna, Erick Gomez-Nieto (2019): Generating Audiovisual Summaries from Literary Works using Emotion Analysis, 32nd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI).
> `https://ieeexplore.ieee.org/document/8919684`

Motivation

The authors present an application of emotion analysis in the area of music generation. Their tool converts the emotions prevalent in a literary work into a musical piece that can accompany the reading process. Moreover, they visualize the interactions between the characters in the book using graph animations but since they are independent of emotions, we will neglect them here. These audiovisual summaries should enhance the reader's experience and provide basic insights about the story.

Data

To recognize the emotions in the text, they make use of the NRC Word-Emotion Association Lexicon which maps words to the eight emotions by Robert Plutchik. Only six of those emotions are kept, namely *joy*, *sadness*, *anger*, *fear*, *surprise*, and *anticipation*. They test their method on five bestseller novels from different genres.

Method

Each novel runs first through a text preprocessing pipeline including the conversion of abbreviations, the removal of stopwords and punctuation, and stemming. A book is then split into four sections, and the number of words of each emotion as well as for the two polarities (positive, negative) is counted in each section. By using these counts in hand-crafted rules, a matching music is generated for each section and the transitions between them. For example, the emotion *fear* should be represented by a dissonant melody. The music generation consists of three parts: 1) the harmony as a progression of chords revoking tension, 2) the rhythm consisting of tempo, time signature and note duration, and 3) the melody which chooses a suitable octave and scale, and assigns musical notes to the rhythmic pattern. Finally, the piece is generated using instruments matching to the associated emotions.

Main Result

Two user studies were conducted to evaluate the method. Their outcome shows that the generated musical pieces mainly represent the intended emotions. Looking in more detail into the result for one novel, the audiovisual summary was judged to be generally matching to the content of the text and useful for its understanding but it was also perceived as not fully assimilating to the novel.

Critical Reflection, Limitations

The paper presents an interesting idea of applying emotion analysis to literature and linking it to music generation. Although their rule-based approach might produce melodies without much data, it might not scale well to other text domains, emotion sets, and music genres, and seems to produce the same melody for the same set of emotions. This could be resolved by a more flexible approach, like a deep learning system. Moreover, it is not clear to which extend the generated musical pieces are suitable for actual applications. For instance, they might not be appropriate or long enough for the use in an audio book. Thus, a more detailed examination of this topic would be necessary.

Nishan Chatterjee

Cowen, Fang, et al. (2019): What music makes us feel: At least 13 dimensions organize subjective experiences associated with music across different cultures, PNAS. `https://doi.org/10.1073/pnas.1910704117`

Motivation

How does music evoke feelings in listeners: Are feelings constructed from general affective features or is it the other way around? How many distinct feelings do we experience in response to music? Which of the feelings associated with music are dependent on the cultural background of the listener, and which are consistent across cultural groups? How are the feelings that music evokes distributed: Discrete or Continuous? These are the questions addressed by this paper.

Data

Participants generated multi-emotion labels for 1841 data points with 28 feelings and 11 broader affective features and rated them on a 9-point Likert scale.(Three groups: {Collecting Library of Music, Choose feelings, and choose broad affective features}. Total participants: 1591 US, 1258 Chinese. This amounted to a total of 375,230 judgments of all music samples. 2 additional sets of music samples were collected to examine the primacy of categories in feelings associated with Western and Traditional Chinese Music (Valence and Arousal).

Method

Cross-cultural similarities in subjective experience was measured by computing the inter-annotator agreement and Monte Carlo simulations. Signal correlations were used capture the degree of similarity to associate feelings and affective features and linear regression to analyze them. Principal preserved component analysis (PPCA) was used to measure the distinct varieties of subjective experiences that were significantly preserved across the US and Chinese judgments. Feelings associated to music were found using factor rotation (varimax) applied to the 13 significant components extracted using PPCA. To find if the feelings were distributed across a gradient of subjective experience, t-distributed stochastic neighbour embeddings(t-SNE) was used. An online interactive map can be found here. The Pearson and the Spearman correlations of the standard deviations showed that there's a blend of categories which aren't entirely miscible. For understanding if music conveys what it's intended to convey vs the primacy of specific feelings being a property of only Western music, a confirmatory approach was used instead of PPCA.

Main Result

Agreement for Energizing and Triumphant music was preserved more than Valence and arousal across the two cultures. Feelings associated with music occupy continuous gradients which overturns a long standing theory of discrete spaces. The semantic content of the videos also help reliably distinguish subjective experience across a single culture.

Critical Reflection, Limitations

t-SNE visualization of the subjective emotion experience is an excellent way to represent insights. The paper addresses how there's definitely more dimensions present as 2549 is too few data points. Groups analyzed are also not homogeneous which makes the resulting insights of each cluster imprecise. The paper doesn't address how subjective experience across cultures may occupy continuous gradients.

Alexander Christodoulou

Issei Fujishiro, Anri Kobayashi: Ambient Music Co-player: Generating Affective Video in Response to Impromptu Music Performance, ITE Transactions on Media Technology and Applications, 2021, Volume 9, Issue 1, Pages 2-12. `https://www.jstage.jst.go.jp/article/mta/9/1/9_2/_pdf/-char/en`

Motivation

Adding to similar research, the authors improve an *affective video-generation system* for musicians which enhances performances by recursively analyzing and evaluating musical input in real time.

Data

The authors record continuous audio signals from an electric guitar connected to an effects unit. The audio signal is captured in time-limited *frames*, which are accumulated to form *performance data*. Performance data consists of arrays of notes, *phrases*, which fulfill musical functions.

Method

The system extracts emotional features from performance data: (1) **sound-level intensity** and **rhythm strength** are translated as arousal, and (2) **timbre** and **rhythm regularity** are translated as valence. Sets of these features are transformed into vectors, and evaluated using SVM-based algorithms trained on hundreds of guitar performances. Output labels correspond to one out of nine possible emotion categories. For visualization, the authors employ a video-generation method based on a 2-D array of *cells* where each cell is dependent on neighboring cells and may dynamically change its appearance. The resulting video is an input-based generated animation of different shapes and colors. Finally, after being tested on a subject, the system performance is evaluated through a self-evaluation questionnaire.

Main Result

The authors infer that the system achieves the primary goal of *inspiring a player*. However, the results don't seem to be comprehensively interpretable given the project scope. Prospectively, the authors are looking to carry out more expressive experiments allowing to isolate the system's affective aspect. Future plans also include adding multimodal measurements like skin conductivity, and expanding the system to interact with groups rather than only individual players.

Critical Reflection, Limitations

Funded by grant-in-aids for **Scientific Research on Innovative Areas** as well as **Challenging Research**, this paper presents itself to tackle a pioneering field of research where comparability is an apparent issue. It tries to combine existing disciplines for which there has been no standardized research framework. Therefore, the authors should establish a multidisciplinary framework, which includes defining and agreeing on standard scores and measurements for typical tasks and experiments. The evaluation and quantification of *a musician's playing performance*, an essentially subjective matter, could be improved by introducing measures going beyond the current self-evaluation questionnaire. Further issues include (1) the number of test subjects, which should be significantly increased in order to solidify the meaningfulness of experiments, as well as (2) the limitation to a single instrument. The authors apper to claim generalizability for musicians in general. However, although this seems intuitively comprehensible, the paper only delivers somewhat tentative results for guitar playing specifically.

Sensors

Xanat Herrera

Liu, Y., Sourina, O., & Nguyen, M. K. (2011). Real-time EEG-based emotion recognition and its applications. In Transactions on computational science XII (pp. 256-277). Springer, Berlin, Heidelberg. https://link.springer.com/chapter/10.1007/978-3-642-22336-5_13

Motivation

The authors' main goal was to enable users to interact with digital media through their emotions in real time. To do so, electroencephalogram (EEG) signals were chosen, since they cannot be controlled or faked by the user, unlike voice or facial expressions.

Data

The data used was manually collected through two experiments, in which EEG signals of 10 to 12 subjects were recorded when playing audio stimuli labeled with valence and arousal values. Afterwards, the subjects filled a questionnaire where they self-assessed their valence and arousal levels, and described their feelings in their own words.

Method

A two-dimensional Arousal-Valence model was chosen for emotion classification. The real-time emotion recognition procedure uses an algorithm to calculate FD values from the EEG data collected through the Emotiv headset. Then, the arousal level is obtained independently by calculating the FD value of one electrode. After that, the difference in FD values between the right and left hemisphere, called lateralization, is used to obtain the valence value. Finally, the combination of arousal and valence values are mapped to its corresponding discrete emotion label.

Through the experiments carried out for data collection, it was found that the threshold for high/low arousal and the laterilization pattern vary across individuals. To tackle this problem, a training session for the emotion recognition system was devised, which uses EEG data with emotion labels of the particular subject as input to obtain personalized thresholds and laterilization patterns.

Main Result

An evaluation of this method was not included in the paper. Using this system, however, the authors implemented three prototype applications. The first allows real-time visualization of emotions as facial expressions on a personalized avatar in a 3D collaborative environment. The second is a music therapy site where the music is automatically adjusted in real-time to have the desired emotional effect on the patient. Lastly, a music player which plays songs corresponding to the user's identified emotion was implemented.

Critical Reflection, Limitations

Regarding the method proposed, the training session would probably be hard to implement in a real-life application, since it requires EEG signals labeled by the user. Nonetheless, I found the proposal very interesting, especially how FD values can be matched to valence and arousal values. Maybe future research could be done to see if FD values can be matched to other dimensions, such as those described in Appraisal Theories, and in that way perhaps achieve a more robust model. I believe the main drawback of this paper, however, is the missing evaluation of the system's performance, and error analysis, which impedes an objective judgement of the method used and the possibility to compare it to future proposals.

Eileen Wemmer

Guillaume Chanel, Cyril Rebetez, Mireille Bétrancourt, Thierry Pun (2011): Emotion assessment from physiological signals for adaptation of game difficulty, IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans. `https://doi.org/10.1109/TSMCA.2011.2116000`

Motivation

This paper aims to answer the question of whether or not controlling the difficulty level of a game based on the player's emotions, as classified based on EEG and peripheral signals, is a viable option to keep them engaged. More specifically, they seek to find out two things. First, does the widely known flow theory hold for video games, i.e. do different difficulty levels relative to a player's skill-level elicit different emotions in that player? And second, are peripheral and EEG signals viable indicators for the current emotion of the player?

Data

The authors conducted a study on 20 people to gather data. They evaluated their current skill level and adapted the speed of falling blocks in a game of Tetris accordingly, introducing an easy, meadium and hard mode for each. They then had each of them complete six gaming sessions, each starting with a short resting period to simultaneously let their physiological signals return to a baseline and record that baseline. This was followed by one of the three gaming modes in a randomized order and a questionnaire. All participants had their peripheral signals recorded, for 13 of them additional EEG data was gathered.

Method

To answer the first question, the authors performed statistical analyses on the questionnaires. For the second question, they evaluated three different classifiers and three different feature-selection algorithms on the gathered data. To approximate real-world use-cases, the trained model needed to be applicable for players it had not been trained for. For each participant, the authors therefore trained each of the models with the data of all the participants except that one. To calibrate the model, they only used one minute worth of physiological signals gathered as a baseline. All this was done for only the peripheral signals, only the EEG signals, and a combination of both.

Main Result

The authors were able to support an affirmation of the first question. This indicates that a player's emotions may generally be good grounds to base a game's difficulty upon. They achieved their maximum emotion-recognition accuracy at 63% by using both peripheral and EEG data.

Critical Reflection, Limitations

The paper does a very good job of finding out whether or not the basic idea of linking players' emotions with a game's difficulty level is viable by considering both the correlation between the two and the technical viability. While the authors go to great lengths to show the learned models don't have to be retrained for each new player to ensure a better evaluation of the approach's applicability, the modalities they used don't necessarily hold up to that same standard. While they suggest fusing the signals with those of further modalities, given the recent advances in Deep Learning, the rise of mobile gaming with it's constant possible access to frontal-camera and microphone data, repeating the experiments now might make an even stronger case for the idea's usage in modern-day games.

Iveta Senkane

> Antonio Fernández-Caballero et.al. (2016): Smart environment architecture for emotion detection and regulation. Journal of Biomedical Informatics 64, 55–73. https://doi.org/10.1016/j.jbi.2016.09.015

Motivation

Generate a complete three-stage pipeline architecture to detect the emotional state of a subject by analysing physiological signals, facial expressions and behaviour using smart sensors in order to assess if the environment in smart health facility should be modified by music and colour/light to guarantee a subject healthy (positive) emotion.

Data

Series of face image files with annotation attached to each; major Database repositories: (a) Informatics and Mathematical Modelling - model of 58 facial points with analysis of 37 images of frontal faces; (b) BioId - 1521 frontal face images, labeled with 20 points; (c) Extended Multi Modal Verification for Teleservices and Security (XM2VTS) - 2360 images with 68 marked-up facial features and real-time videos and physiological signals. Approximately 30 healthy volunteering elderly participants.

Method

Pipeline: (1) Emotion Detection (facial emotion detection, behaviour detection and Valence/Arousal detection); (2) Emotion Regulation (music and colour/light as pleasant/unpleasant) and (3) Emotion Feedback control (multimodal fission analyses the results from Emotion detection layer and provides the commands by adjusting them to the Emotion regulation layer).

INT3-Horus multi-sensory framework is used for monitoring and activity interpretation; ASM tracks the fiducial points coarsely (geometrical features/distance - eyebrows, eyes, mouth and nose) to classify one of 6 basic emotions (Ekman) plus Neutral; and SVM - to generate classification model operating in real-time. Larsen and Diener's circumplex model of affect (CMA) enriched with 6 basic emotions as in Russel's CMA allows Behaviour Detection by label Active or Inactive and Valence(pleasantness)/Arousal(activation) Detection and classification with ANN as excited or nervous and relaxed or bored. The Behaviour Detection from body image is realized by subjects detection by camera (colour level/ shadow detection algorithm). The Valence/Arousal Detection from physiological data is proceeded by EDRS and a HRS which provides how pleasant/unpleasant and how exciting/calming a stimulus is.

Main Result

The architecture allows to track a person among the other people in a room, however does not allow to build a reliable system to quantify to which extend the proposal is capable to achieve the goal.

Critical Reflection, Limitations

This proof-of-concept offers a better understanding of how the technical aspects may be combined, however the multidisciplinary approach seems to be way too complex and demands even more fine-grained procedure and time to succeed at emotion regulation. Besides, the private data has to be also handled according to the ethical rules. This paper was published shortly after completing the first phase of the project, therefore it is difficult to say, if the idea is realizable any soon.

Robotics

Carlotta Quensel

J. C. Kim, P. Azzi, M. Jeon, A. M. Howard and C. H. Park (2017). "Audio-based emotion estimation for interactive robotic therapy for children with autism spectrum disorder", *14th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)*, Jeju, pp. 39-44, `https://ieeexplore.ieee.org/document/7992881`

Motivation

Therapy robots can help children with autism spectrum disorder to recognize emotions by splitting emotional reactions onto a facial and gestural robot. To generate appropriate reactions and interact directly with children, real time emotion analysis from speech is needed.

Data

To train the emotion classifier, the IEMOCAP (Interactive Emotional Dyadic Motion Capture) database was used. The database contains 12 hours of dialogue between five female and five male actors. The sessions are about 5 minutes long and every phrase is annotated by three human evaluators with valence, dominance and activation scores between 1 and 5. These scores map onto the categorical emotion labels excitement, frustration, disgust, fear, surprise, anger, sadness, happiness, and neutral state.

Method

The speech features were extracted from the IEMOCAP using openSMILE and projected onto 150 principal components for dimension reduction. For the real-time implementation, the dimension reduced features are normalized using only a few 'neutral state' data samples for each speaker. Thus, the algorithm can estimate the sound of a person in neutral state and classify emotions from there. The classifier was trained as three SVMs each for one dimension of the VAD score with a linear kernel function. For a new speaker, the normalization is performed in the same manner and the features are passed to the three SVMs.

Main Result

The new method of normalizing speakers improves the classifier compared to previous methods. The classifier has an unweighted accuracy of 62.9% for arousal and 52.7% for valence, which a significant improvement compared to general speaker normalization which is a promising result for real-time classification.

Critical Reflection, Limitations

While the accuracy for valence and activation is acceptable, dominance shows low scores, which is a frequent problem in speech analysis. Because the classifier has to work in real-time, the method does not use a multi-temporal approach, which would improve accuracy but slow down the classifier. The paper evaluates the classifier on data from neurotypical adults, therefore to use the findings in therapy robots for autistic children, more speech data from (neurodiverse) children is needed, as well as general evaluations on child-robot interactions.

Felix Bühler

> Jelle Saldien, et al. (2010): Expressing Emotions with the Social Robot Probo, International Journal of Social Robotics. `https://link.springer.com/content/pdf/10.1007/s12369-010-0067-6.pdf`

Motivation

A study has shown that only 7% of affective information is transferred by spoken language 38% is transferred by paralanguage, and 55% of transfer is due to facial expressions. A social robot Probo was developed to do further research on the recognition of emotions.It has its own identity, including a name, a history. Animal assisted therapy/-activities are expected to have useful psychological, physiological, and social effects on people. For several reasons attempts are made to replace these animals by robots.

Prototype

Probo is a huggable robot and looks like an animal. Probo is equipped with animated ears, eyes, eyebrows, eyelids, mouth, neck, and an interactive belly-screen and a trunk to intensify certain emotional expressions. The robot has 20 degrees of freedom in its head. The robot is covered with a foam layer and a removable fur-jacket. For online-evaluation, a 3D virtual model of Probo has been created. To represent the emotions, the Facial Action Coding System (FACS) (by Ekman and Friesen)was used to define for each emotion-specific Action Units (AU), which correspond to muscles in the face. The basic facial expressions are represented as a vector in the 2-dimensional emotion space based on Russel's circumplex model of affect (valence-arousal). To have a smooth transition between emotions, the current emotion could always be interpolated with the neutral emotion, which has to be used when switching.

Method

The first pilot study, the virtual model was used to test the recognition of facial expressions. For other tests, Probo was used uncovered or covered. A questionnaire displaying pictures and a multiple-choice of emotion words were used. The same amount of children and adults took part.

Main Result

For the evaluation of the virtual prototype adults and children were to classify 8 emotions (Eckman + *neutral* + *tired*). Adults had an accuracy of 67% and children 60%. Further evaluation for the Eckman emotions has been done with children. For the virtual prototype, they reached an accuracy of 88%, the uncovered 83%, and for the covered one 84%. There has been no significant difference found based on the gender or age of participants.

Critical Reflection, Limitations

The initial development was done with the virtual avatar (resulting in good accuracys). Because the fur behaves differently than in the virtual world, the final evaluation-accuracy (with the mechanical prototype) differs. The mouth e.g. is very hard to recognize because it is covered compared to the uncovered/virtual prototype. Without the trunk some emotions would be very hard to be recognized e.g. disgust and sadness have very subtle changes. Adding more activation units could solve this problem. The prototype was only evaluated for facial emotions, but body-language like gestures are completely ignored. This would further improve the recognition. The robot has already implemented several input types, but not automatic process has been done so far.

Maximilian Wegge

C. P. Lee-Johnson and D. A. Carnegie (2010): 'Mobile Robot Navigation Modulated by Artificial Emotions', *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 40, no. 2, pp. 469-480.
`https://ieeexplore.ieee.org/abstract/document/5282573`

Motivation

In the field of robotics, research on emotions primarily focuses on human-robot interaction (social robots). However, modelling emotions for 'non-social' robotic tasks has not been widely considered yet, even though biological evidence suggests that emotions are beneficial to coping with unfamiliar situations. This paper presents an approach to model emotions in a manoeuvrable robot, which increases its capability to navigate through a three-dimensional space.

Data

The robot collects data about its surroundings with multiple sensors which provide it with the data required to accomplish the navigating task. The quantitative and qualitative data relevant for evaluation (the robot's movements, time elapsed until goal is reached, amount of collisions) is obtained by conducting the navigation tasks in a simulated setting.

Method

The approach considers the emotions *fear* (reduce speed when estimated probability for collision exceeds threshold), *anger* (reduce obstacle aversion when progression stagnates due to external obstructions), *surprise* (explore environment before continuing towards goal when sensor feedback contradicts the previously generated map), *sadness* (neg. reinforcement when behaviour does not result in reaching the goal) and *happiness* (pos. reinforcement when behaviour results in reaching the goal). The robot's performance is tested in several environments (hardcoded as well as procedurally generated) that each impose different challenges. Multiple attempts per environment allow the robot to learn about the given challenges and adapt to them. Various settings are considered: each emotion activated individually as well as all five emotions activated at once. The results are evaluated against the performance of a 'non-emotional' robot in the same environments.

Main Result

Due to its 'emotional' behaviour, the robot experiences lesser collisions and reaches the goal in a shorter amount of time than the 'non-emotional' baseline approach. For example, *fear* allows the robot to navigate at an overall higher speed while only reducing its speed when close to obstacles (improvement over static speed); *anger* enables the robot to break through (breakable) obstacles when no other path is available. However, *Fear* has a negative side-effect: The robot becomes 'afraid' of a narrow passage and will not traverse it. This is eventually resolved when *anger*, triggered by the persistent obstruction, suppresses the *fear*.

Critical Reflection, Limitations

Although unconventional and different from other approaches of modelling emotions, the approach still reflects the opposition of emotions in Plutchik's wheel. However, the interpretation of *surprise* is not entirely appropriate. Intuitively, *surprise* would result in the re-evaluation of the current situation and goals instead of prioritizing exploration (which is more similar to 'curiosity'). Further research could investigate if this approach is of similar benefit for other robotic applications.

Personality

Hanna Surjadi

> Barbara Plank, Dirk Hovy 1 (2015) : Personality Traits on Twitter—or—How to Get 1,500 Personality Tests in a Week,Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis `https://www.aclweb.org/anthology/W15-2913/`

Motivation

Plank and Hovy are taking a data-driven approach for identifying personality traits, based on the self-identification of twitter users. By doing so they can test their model on a large set of data. Which is how this papers approach is different compared to previous works. The authors try to differentiate between the four binary dimensions of MBTI. (Introverted - Extroverted, Intuitive - Sensing, Thinking - Feeling, Judging - Perceiving)

Data

The data is gathered by searching mentions of the Personality Type Indicators and "Briggs" on Twitter. Any tweets that contain more than one Indicator are manually removed. Additionally, the gender and other count-based meta-features are labelled. In the end the data contains information about 1500 distinct users. Comparing the general population estimates the corpus is shifting towards females and introverts. Which supports findings about the correlation between online-offline choices of introverts and extroverts..

Method

The authors train a logistic regression classifier. The features are n-grams, gender, and several discretized count-based meta-features. Stopwords are not removed because they have shown to harm performance. Pre-processing consists of tokenizing and replacing hashtags, URLs, and usernames with unique tokens.

Main Result

Compared to a majority-class baseline, the model improves on the I-E and F-T dimensions. However, there is no improvement on S-N and even a drop in P-J. The gender-controlled dataset does not show significant differences. Linguistic features are the most predictive features for personality, but meta-features such as follower count or status count also support findings. Gender seems to be an effective feature in predicting the F-T dimension, but less effective for distinguishing I-E. The S-N and P-J dimensions refer to a person's perception. The model does not learn how to predict them, which indicates that either there is no linguistic evidence for them, or other features are needed..

Critical Reflection, Limitations

The study relies on the self-typing by Twitter users. Besides the fact that MBTI is not scientifically valid, most MBTI tests that are available are automated and not very thorough. Despite this the paper depicts a pragmatic approach, which might be still valuable seeing that there is no annotated corpus on personality traits that can be compared in size. Barbuto (1997) suggested a reconstruction of the MBTI. However, all relating papers that were cross read (in a short timespan), still use the binary four dimensions of MBTI instead of using an approach that depicts the differences between all eight Jungian cognitive functions. I would like to see a version of Planks and Hovys approach that tries to prove or disprove linguistic features of the cognitive functions.

Hasan Cengiz Öztürk

Lin Qiu, Han Lin, Jonathan Ramsay, Fang Yang(2012): You are what you tweet: Personality expression and perception on Twitter, Journal of Research in Personality http://www.sciencedirect.com/science/article/pii/S009265661200133X

Motivation

In this paper, microblogging services, such as Twitter, are examined concerning consumers' language and resulting personality traits. This linguistic research is conducted to better understand the relationship between personality traits and microblogging through communication. This paper's researchers claim that microblogging conversations are in a "naturalistic" setting within such services and are critical to understanding the actual thoughts and feelings of the people who are posting them. The paper argues that it is much more important to analyze such an everyday environment than to only look at well-controlled surveys in "decontextualized environments."

Data

This paper's background is how tools like LIWC process text information and assign personality traits to linguistic patterns in writing styles and self-reported personality traits. According to zero acquaintance research, personality judgment was also possible with these linguistic tools that would look at how people would name themselves in online games and communicate via email. Brunswick's lens model was used as a base to look at the relationship between microblogging, personality, and interpersonal perception.

Method

This paper's background is how tools like LIWC process text information and assign personality traits to linguistic patterns in writing styles and self-reported personality traits. According to zero acquaintance research, personality judgment was also possible with these linguistic tools that would look at how people would name themselves in online games and communicate via email. Brunswick's lens model was used as a base to look at the relationship between microblogging, personality, and interpersonal perception.

Main Result

In their analysis, Qiu et al. focus on inter-observer consensus and accuracy, cue validity, cue utilization, sensitivity, the mediating role of linguistics cues in self-other agreement, and finally, differences in gender, age, and ethnicity. The results mostly reflect the consistency of personality-related behaviors across different contexts with a few exceptions. Sensitivity was calculated via vector correlations proposed by Funder and Sneed resulting in a consistent result regarding zero acquaintance research. Mediation analysis was done using Preacher and Hayes' INDIRECT macro resulting in mediation on accuracy, especially regarding neuroticism. Some words that were used were not closely correlated to personality but could be better linked to ethnicity, gender, and age. A test was done with a procedure by Giliath et al. and resulted that the testers relied more on linguistic cues rather than stereotypes regarding gender, ethnicity, and age.

Critical Reflection, Limitations

Accurate judgment by unknown testers can be made about the personality traits of microbloggers using linguistic cues. Past research has common results and can be compared to texts in microblogging services strengthening the paper's argument. Novel associations could also be observed, such as the positive correlation between extraversion and assent words. The findings of this paper also have significant theoretical (zero-acquaintance theory regarding microblogging), practical (Twitter becoming increasingly popular), and commercial (psychological profile analysis of the user) use.

Ema-Maria Zaviacicova

Mohammad, Saif & Kiritchenko, Svetlana. (2013). Using Nuances of Emotion to Identify Personality. AAAI Workshop - Technical Report.
https://www.researchgate.net/publication/256437701_Using_Nuances_of_Emotion_to_Identify_Personality

Motivation:
Following up on the research about how lexical categories (e.g., pronouns, tenses and sentiment) have significant correlations with personality traits, this article would like to confirm that emotional categories, particularly fine emotions (e.g., excitement, guilt, yearning) are significant indicators for personality traits.

Data:
For this experiment, they drew from three large lexicons. The NRC Hashtag Emotion lexicon (fine-grained affect categories), Osgood dimension lexicon (coarse affect categories), and the specificity lexicon using Pedersen's precomputed scores of WordNet synset specificities. The corpus for the analysis was an Essay dataset collected by Pennebaker and King (1999). It consists of 2469 essays written by psychology students.

Method:
The NRC Hashtag Emotion Lexicon was used to identify a degree of association between the word and an emotion (called the PMI). Here a large score represents stronger associations. For each emotion-related hashtag, a separate feature (overall 585) was created. For comparison the simple NRC Emotion Lexicon with eight basic emotions, creating eight features, was used. The Osgood dimension lexicon was used for calculating the average evaluation efficiency of the words in an essay. A further question was if people of a certain personality type tend to use terms with high specificity. For identifying each of the five personality dimensions, five Support Vector Machine classifiers were trained. Each result was averaged over three-fold stratified cross validation. LibSVM with linear kernel and six feature groups (Mairesse Baseline, Token Unigrams, Average Information Content, Coarse affect category features, basic emotion category feature, fine emotion category feature) were used. The results were compared with gold labels for yes or no for each of the five personality dimensions. The experiment was repeated ten times, and the scores were compared with a paired t-test.

Main Result:
The fine-grained emotion features from the Hashtag Lexicon offered a statistically significant improvement over the Mairesse baseline, whereas the coarse affect features from the NRC lexicon failed to provide significant improvement.

Critical Reflection, Limitations:
Overall this paper confirmed the assumptions which formed at the beginning of the reading. The information we get from statements with fine grained emotions is more specific and better suited to assign to specific personality traits. On the other hand, it is more difficult to get this information from coarse emotions. For further analysis, it would be interesting how effective this assignment to personality traits is in other more common corpora like e-mails or tweets, and how we could use this knowledge for, e.g., improving customer service.

Health

Leonie Lanfrit

Hamed Khanpour, Cornelia Caragea (2018): Fine-Grained Emotion Detection in Health-Related Online Posts, Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. https://www.aclweb.org/anthology/D18-1147

Motivation

In Online health communities (OHC) patients and their families and friends share their thoughts about topics like mental health, therapies and side effects. It would be very useful to know what kind of emotion a patient is having, so that caretakers or doctor can offer improved help. The goal of this paper is thus to automatically detect fine-grained emotion types from health-related posts with a computational model that is able to exploit semantic information from text.

Data

The authors construct two health-related datasets. The first is created by using data from Biyani et al. (2014) and contains 1066 sentences from the breast cancer discussion board in the Cancer Survivors' Network (CSN). They build the second dataset of 1041 sentences out of comments from the lung cancer discussion board of CSN. The multi-label annotation of the corpora follows the six emotions suggested by Ekman (1992), and the annotators met with researchers to discuss disagreements to finally achieve 100% Kappa inter-annotator agreement. 39,7% of the annotated sentences were labeled with *joy* and 15% with *sadness*. No other emotion achieved the 5% mark.

Method

They build a computational model that combines the output of a Convolutional Neural Network (CNN) with lexicon-based features, which are then fed into a Long Short-Term Memory network and call it ConvLexLSTM. In addition to two emotion lexicons, they use the same lexicons as Biyani et al. (2014) that include cancer drugs, side-effects and therapeutic procedures. They construct variants of their model (e.g without lexicons, only CNN) and train and evaluate their models in a two-class setting only on the emotions *joy* and *sadness*, because they have at least 5% coverage in the labeled data. Finally, they apply their model in a large scale experiment and compare the emotions *joy* and *sadness* of users on the CSN two days before and after and on the actual holiday.

Main Result

ConvLexLSTM achieves the best results throughout all experiments, thus all components contribute to the detection of emotions. The importance of capturing high-level semantic information for emotion detection via deep neural networks is shown by worse results with Support Vector Machines. They then find out that there is an increase of *joy* and a decrease of *sadness* on Thanksgiving and Christmas, possibly due to many social events. On New Years Eve the percentage of *sadness* is twice as high as *joy*, maybe because of the end of the holiday season.

Critical Reflection, Limitations

I liked that the paper is easily understandable and has a good structure. The emotional state can indeed be an important indicator of what patients need in their therapy, so this task can make their lives better. It would be interesting to see the results if they trained and evaluated their model with multi-classes and with all six emotions, as the results will probably not be as high anymore. Another question that arises is how they chose the annotators and if they were experts in the field. And finally, I'm wondering what exactly do holidays have to do with people that are sick?

Rajesh Baidya

> Munmun De Choudhury, Michael Gamon, ... (2013): Predicting Depression via Social Media, International AAAI Conference on Weblogs and Social Media . `https://ojs.aaai.org/index.php/ICWSM/article/view/14432/14281`

Motivation

Collect assessments from Twitter users who report that they have been diagnosed with clinical MDD (Major Depressive Disorder), introduce several measures and use them to quantify an individual's social media behavior for a year in advance of their reported onset of depression. And Ultimately to build a classifier that can predict whether an individual is vulnerable to depression or not.

Data

Total 2,157,992 posts from 1 year of Twitter posts from a set of 476 users contained 243 males and 233 females, with a median age of 25. CES-D questionnaires used as the primary depression level estimation tool. Two classes of users: an MDD positive class of 171 users (36%), and a negative class of 305 users.

Method

Measures:

- **Engagement:** Volume, insomnia index, etc.
- **Emotion:** 4 emotional state of users. positive affect (PA), negative affect(NA),activation, and dominance. Psycholinguistic resource LIWC and ANEW lexicons are used to determine states.
- **Linguistic Style** 22 specific linguistic styles, e.g.: articles, auxiliary verbs, conjunctions, etc.
- **Egocentric Social Graph, Depression Language**

Classification:

- Time-series measure per user was constructed over the entire one year of Twitter history. Feature vectors were extracted from this Time-series data.
- **Feature Vectors:** Total 188 features (there are 43 dynamic features, 4 demographic features).
- Support Vector Machine classifier with a radial-basis function(RBF) kernel for classification

Main Result

Individuals with depression show lowered social activity, greater negative emotion, high self-attentional focus, increased relational and medicinal concerns, and heightened expression of religious thoughts. Classifier yielded promising results with 70% accuracy.

Critical Reflection, Limitations

This is a very old Paper. With the implementation of the latest machine learning technique results can be improved.

Van Hoang

> Lina M. Rojas-Barahona, Bo-Hsiang Tseng, Yinpei Dai, Clare Mansfield, Osman Ramadan, Stefan Ultes, Michael Crawford, Milica Gašić ... (2018): Deep learning for language understanding of mental health concepts derived from Cognitive Behavioural Therapy, Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis. `https://www.aclweb.org/anthology/W18-5606/`

Motivation

The authors aim to train a model which is capable to understand mental health concepts derived from Cognitive Behavioural Therapy (CBT). This model is hypothesized to be incorporated into a dialogue system which can deliver therapy.

Data

The corpus contains 500k written posts on Koko platform[1]. This platform is for those who are being stressful and having negative thoughts. They can post anonymously there and other users can read and offer consolation. From these posts, a total number of 4035 posts is to be analyzed and annotated by two psychological therapists. Each post is labelled with *thinking errors*, *emotions*, and *situations*. There are 15 *thinking errors*, 9 *emotions*, and 7 *situations* to choose from. These are analyzed from 1000 posts. The annotators can pick as many labels for each category as they deem fit. In other words, this is a multi-label corpus. Kappa scores for two annotators are 0.61, 0.90, and 0.92 for *thinking errors*, *emotions*, and *situations* respectively.

Method

The problem of learning CBT concepts is cast as classification problem. Two DL models a CNN and a GRU, and two ML models, LR and SVM, are experimented with. They also tried with different pre-trained embeddings, namely GloVe, Skip-Thought, and Doc2Vec. For LR and SVM, bag-of-words method is used.

Main Result

Results are reported in averaged and weighted averaged F1 scores. Best performance is obtained on CNN-GloVe models with F1 scores of 0.454, 0.612, and 0.578 for *emotions*, *situations* and *thinking errors* respectively. However, the difference in scores for the labels in all categories is very high, ranging from 0.8 (best score) to 0.15 (lowest).

Critical Reflection, Limitations

Though the authors stated that *thinking errors* is a difficult classification task due to low inter-annotator agreement, the results show that its scores are actually higher than that of *emotions*. They suggested that future direction could include richer distributed representation and extend the current category and its labels with more concepts. We, on the other hand, don't think that extending the number of labels/concepts is a good idea. From their reported scores, it seems to us that low scores on some labels are due to low number of training data for these particular labels. For example, in *emotions* category, label *Anxiety* has 2547 samples while *Jealousy* only has 126 samples. Therefore, we believe that one direction is to try to make the labels *less* fine-grained.

[1] `https://www.koko.ai/`

Nina Doerr

> Denecke, Kerstin; Vaaheesan, Sayan; Arulnathan, Aaganya (2020): A Mental Health Chatbot for Regulating Emotions (SERMO) - Concept and Usability Test, IEEE `https://ieeexplore.ieee.org/document/9000924`

Motivation

Pharmacotherapy and psychotherapy are the established methods to treat mental disorders like depression or anxiety. Today there is gap between between available professionals and people affected by mental disorders. As a consequence some researcher developed machines that can support patients in self-help when waiting for a therapeutic session or between sessions. Previous work is still limited in clinical evidence and solely focuses on English language. Therefore Denecke et al. present a chatbot for regulating and self-monitoring of emotions for people with mental disorders in German language.

Data

Denecke et al. aggregated data from interviews and literature search to specify the requirements for their chatbot application. Their application saves data from the following functionalities: the users daily mood as an emotion diary, emotion recognition, supporting measurements and activities, as well as further information about the app and the applied methods. For emotion recognition they used the Emotional Dictionaries of SentiWS, which links German words to emotions.

Method

In general the application is grounded on the cognitive behavioural therapy (CBT), a method for treating depression and other mental disorders with focus on emotions. Their mobile application simulates a realistic conversation by giving appropriate answers to user input messages. For that, the chatbot identifies the underlying emotions (fear, anger, sadness, joy and grief) in the messages with a lexicon-based approach. The used emotions are described in Richards Graph C-I-E Theory and are also present in the used lexicon. The emotion recognition and analysis is based on the ABC theory by Albert Ellis to find the trigger of a specific mood. The implemented algorithm first splits the input message into sentences, tokenizes them, removes stop words (e.g. prepositions) and then classifies them into categories. If the chatbot can not find a specific emotion, it asks the user to select one. After that the chatbot proceeds based on the identified emotion.

Main Result

For the evaluation of the user experience and the quality of the chatbot application, Denecke et al. conducted a usability test with 21 participants. It showed that efficiency and perspicuity are good but some participants perceived it as not very stimulating and motivating. They also asked psychologists and psychotherapists for feedback. They responded positively and suggested that a chatbot application could be beneficial for a gap between therapeutic sessions. However they perceived an alerting system could be helpful, if the chatbot detects a certain risk for a patient.

Critical Reflection, Limitations

Denecke et al. state several own improvements that i also would suggest e.g., the missing negation detection. In addition a machine learning approach could be helpful to improve the answers and maybe the detection of metaphors. However the research on mental health applications is very important, since the need or awareness of a good mental health is getting more attention today.

Speech

Pavan Mandava

Han K et al. (2014): Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine, INTERSPEECH-2014. `https://www.isca-speech.org/archive/interspeech_2014/i14_0223.html`

Motivation

This paper utilizes deep neural networks to extract high level features from raw data and show that they are effective for speech emotion recognition.

Data

- Makes use of the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database for training and evaluating the model.
- The database contains audiovisual data from 10 actors, and this paper only makes use of audio data.
- Utterances with 5 emotions: {*excitement, frustration, happiness, neutral, surprise*}

Method

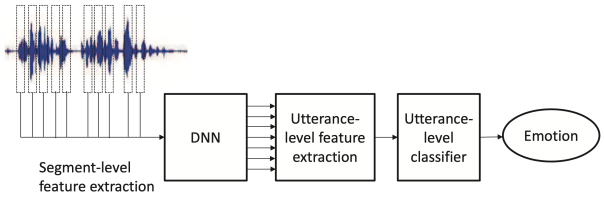


Figure 1: Overview of the Model

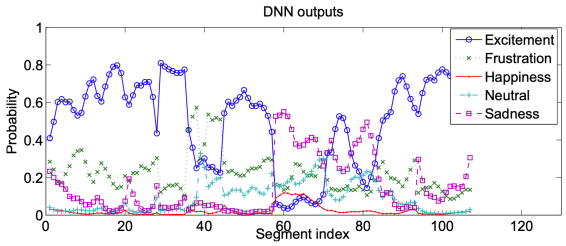


Figure 2: DNN Outputs of an utterance

This approach divides the speech signal into segments and then extract signal-level features to train a DNN. The DNN outputs the emotion state probability distribution(softmax) for each segment. From these segment-level emotion state distributions, utterance-level features are constructed and fed into an **E**xtreme **L**earning **M**achine to predict the emotion of the whole utterance.

- Segment-level features → *pitch-based, energy-based, MFCC features*
- DNN → *Feed-forward Neural Network, 3 hidden layers - ReLU, Softmax output layer*
- Utterance-level features → statistics on probabilities: *max, min, mean, percentages*
- ELM (Utterance-level classifier) → *Extreme Learning Machine, 1 hidden layer*

Main Result

The DNN based approaches significantly outperformed with 20% relative accuracy improvement (both unweighted & weighted accuracy) compared to the state-of-the-art approaches, like Hidden Markov models(HMM) and OpenEAR. It's also worth noting that the kernel ELMs perform slightly better than the ordinary ELMs and the training time of ELMs is much faster than that of SVMs with even better results.

Critical Reflection, Limitations

This approach of extracting high-level features using Feedforward Neural Networks or Multi Layer Perceptron seems to represent the speech data well, it would be interesting to experiment with Convolutional Neural Networks (CNNs) on this task. Increasing the complexity of the Utterance-level classifier (ELM) combined with more features (like *mean* of 10-segment windows) or experimenting with other Neural Network architectures could be an interesting next step.

Victoria Punstel

A. Tawari and M. Manubhai Trivedi, (2010): Speech Emotion Analysis: Exploring the Role of Context. IEEE, vol. 12, no. 6. https://ieeexplore.ieee.org/document/5571815

Motivation

This paper aims to identify important prosodic and spectral features for speech emotion analysis and understand the effects of context on two different databases.

Data

EMO-DB (Berlin Database of Emotional Speech): This database is comprised of emotional speech data from 10 professional German actors (Female = 5, Male = 5) The actors recorded 10 sentences with neutral content in the seven different emotions - {*fear, disgust, joy, boredom, neutral, sadness, anger*}

CVRRCar-AVDB (Audio-Visual Affect Database): This database is comprised of emotional speech data - {*positive, negative, neutral*} from four participants (Female = 2, Male = 2). The database contains audiovisual data from stationary and moving car settings. In this study, only data from stationary car settings were selected.

Method

Prosodic and spectral features were chosen to model the emotional states. A CFSSubset Eval algorithm paired with a stratified tenfold cross-validation procedure was used for feature selection. For each experiment, an SVM trained with an SMO algorithm with 2-10 aggregate features was used. The user-text-dependent context results were developed using tenfold cross validation, while the user-text-independent context utilized a leave-one-subject out cross validation. Finally, the gender dependent context results were based on a randomized tenfold cross validation, leave-one-subject out and leave-one-text out cross validation.

Main Result

A high emotion recognition rate of 84% was achieved for the EMO-DB database, and 87% for the CVRRCar-ADVB. The user-text-independent analysis demonstrated the classifier may be learning the manner in which an emotion is expressed, while ignoring the verbal content to some degree. This suggests a need for speaker adaptation. Incorporating gender information improved the emotion recognition rate by 3% for both databases.

Critical Reflection, Limitations

The development of the feature set based on pitch and intensity contours proved to be somewhat promising. It would be interesting to experiment further on other prosodic features such as speech rate. In the future, it could be more beneficial to incorporate additional speaker information, such as age, and make use of an automatic gender detection system.

Visual/Body

Alina Braitmaier

Sicheng Zhao, Hongxun Yao, Yue Gao, Rongrong Ji, Wenlong Xie, Xiaolei Jiang, Tat-Seng Chua (2016): Predicting Personalized Emotion Perceptions of Social Images, MM '16: Proceedings of the 24th ACM international conference on Multimedia. `https://dl.acm.org/doi/abs/10.1145/2964284.2964289`

Motivation

Traditional methods of emotion prediction consider the dominant emotion of all viewers, but image viewing, especially on social media, is more complex than that and influenced by social, cultural and educational backgrounds. The paper therefore examines personal emotion perception (PEP) based on several factors added to the visual features.

Data

There was no public data-set for PEP yet that could provide benchmarks. This means the researchers accumulated a large image data-set of personalized emotions, in which emotions are represented in VAD-dimensions as well as in a categorical model of 8 emotions and two sentiments. Images, their meta-data and user meta-data were taken from Flickr. Through analyzing title, tags and description, they evaluated the expected emotion of the uploader, through analyzing the comments, they evaluated the actual emotion of the individual viewer.

Method

They considered factors such as visual content, social context, temporal evolution and location influence and combined these factors through rolling multi-task hypergraph learning. To evaluate the influence of each factor, they rotated through multiple executions, leaving a different factor out each time.

Main Result

The proposed method outperforms state-of-the-art approaches. The performance improvement is higher on negative emotions, which implies that the different factors play a bigger role in personalized perception for negative emotions. Stronger social connections also seem to have a bigger influence on performance improvement.

Critical Reflection, Limitations

This paper is concerned with a very current and inevitable research topic. I believe their approach to be innovative and smart, as image perception is in fact greatly influenced by many other factors. So far, the research of this paper is very limited to social platforms, and much of their data can only be collected in that context. I am not sure how or if the application could be expanded to outside of the social media context. Therefore, the method as well as the results should be tested on other social platforms (Flickr differs greatly from Instagram or Facebook), but is very promising for further research into emotion and social dynamics on social media.

Jasvinder Singh

Caridakis et al. (2007): Multimodal emotion recognition from expressive faces, body gestures, IFIP.

Motivation

Present a multi-modal approach which fuses features (facial expressions, body movements, gestures and speech) for recognizing the eight emotions - anger, despair, interest, pleasure, sadness, irritation, joy and pride.

Data

The actors were tasked with recording their gestures, facial expressions and speech patterns based on the emotions. Two DV cameras recorded the actors. One camera focused on the face while the other on the body. Voice recorders were used to record speech samples of the actors. The procedures followed the specifications of the GEMEP corpus. Features were further selected as follows -

- Authors used MPEG-4 FAPs for combining/mapping each facial feature mask. The features were extracted by segmenting the face.
- Body and hands were tracked using the EyeWeb platform which were further extracted based on the emotional cues. The features also correspond to the dynamics of movements which resulted in each gesture comprising 80 motion features.
- A full set of 377 features were extracted from the speech.

Method

A separate Bayes classifier was used for each modality individually. For fusing facial features, feature level fusion and decision level fusion approaches were adapted. The former involved selection of emotion that received best probability in the three modalities. The second approach comes into play when the first approach could not ascertain majority and therefore requiring a voting procedure.

Main Result

For facial expressions, the overall performance of the classifier was 48.3%. There were some misclassifications of emotions with each of it scoring 20%. In the case of emotion recognition for gestures, the overall performance was 67.1% and finally, recognition in the speech modality had an overall performance of 57.1%. The performance of multi-modal classification was 78.3%. This is higher than all of the uni-modal systems. Finally, for a decision based approach, the accuracy score was 74.6

Critical Reflection, Limitations

Well performed experiments. I also liked how they came up with their own pseudo-linguistic language to record speech patterns since actors came from different backgrounds. The paper makes use of more traditional machine learning approaches given that it was published in 2007. The dataset is also not very large which might limit its usability with the current deep learning approaches.

Meghdut Sengupta

> Tadas Baltrusaitis, Daniel McDu, Ntombikayise Banda, Marwa Mahmoud, Rana el Kaliouby, Peter Robinson and Rosalind Picard . . . (2011): Real-time inference of mental states from facial expressions and upper body gestures, Institute of Electrical and Electronics Engineers. `https://dspace.mit.edu/bitstream/handle/1721.1/67458/FERA%20201_McDuff.pdf?sequence=1&isAllowed=y`

Motivation

The central research question of the aforementioned paper is to infer mental states from facial and upper body gestures. They explore a spectrum of feature engineering in terms of facial action units and upper body (shoulder) gestures and perform multilevel Bayesian modelling to detect the mental state (emotion). Emotion set for detection: Anger, Fear, Joy, Relief and Sadness.

Data

The training and test data sets are the GEMEP-FERA datasets, which are the subsets of the GEMEP corpus. The training dataset consists of recordings (average video length of 2.67 seconds) of 10 actors displaying a range of expressions (87 videos). The test dataset (71 videos) contains of six subjects: three of them are present in the training data, while the other three are new. The training dataset did not contain any neutral expressions which they generated as a separate task.

Method

Their approach combines a three step geometric analysis of face and shoulder based features. Firstly, they identified facial action units (AUs) from 22 feature points (found via FaceTracker) and Gabor filters (used for texture analysis). They observed that the corners of eyes and nose tips are static facial features, i.e. the features which remain stable in all kinds of expressions, and modelled neutral expressions based on the same, where they performed hand written action units based on angle and distance computation between the points. Secondly, they used Gabor features to train a binary SVM (one per AU). Thirdly, they used probabilistic models to classify sequences of actions into head and shoulder gestures. They use a 3-symbol HMM for head nods (head up, head down and no action). After detecting edges of shoulders in each frame, they use a line transform method to compute angles of the shoulders from the X-axis, followed by calculating an average angle for right and left shoulders. The angle differences (initial frame - each frame) determine shoulder up, shoulder down and shoulder shake when compared to a threshold. On the the last level they model multilevel Bayesian Network to detect the mental state.

Main Result

Lower face AUs performed worse than the upper face ones. Feature point based system generalized better than the Gabor based system. Basic emotions lack such dynamics, and therefore the system was not as good for their recognition. Hence their system performed better on subtle expressions of relief (F1: 0.538) and sadness(F1: 0.640) and poorly on anger(F1: 0.111).

Critical Reflection, Limitations

Detailed analysis of face and body gestures (Smile, Mouth Open, brow lowerer, head pitch up, upper lip raiser, lips part and cheek raiser) and careful explanation of methodology. Cons: Insufficient training data. The dataset was skewed towards subtle emotions. Further work can focus on analysing hand movements.

Michael Göggelmann

U.M Prakash, Pratyush, Pranshu Dixit, Anamay Kumar Ojha (2019): Emotion Analysis Using Image Processing, International Journal of Recent Technology and Engineering. https://www.ijrte.org/wp-content/uploads/papers/v7i5s2/ES2043017519.pdf

Motivation

Since emotion analysis is still difficult to be applied to images in general, the authors try to find out whether and if, how well "Convolutional Neural Networks" (CNN) suit this task. As a secondary hypothesis they assume that the results could be used meaningfully in a subsequent stress management project.

Data

500 - 600 images from Google images and research datasets from other papers were used as training data. All of these images were labelled 'according to their types'. Afterwards they tested their model on camera snapped images (facial expressions).

Method

The authors used an approach mainly including "Convolutional Neural Networks" (CNN). Their model takes an image as input and after going through certain layers of CNN one emotion out of anger, fear, happiness and sadness gets assigned to the image.

Main Result

The method fits the task of emotion analysis with images as can be seen from a accuracy of 85,23 percent and, moreover, the method is easy to use due to the lack of a need for complex preprocessing. Therefore, the trained model is to be used in a larger project on the subject of stress management in the future.

Critical Reflection, Limitations

The method seems to work just fine, but the evaluation of the results is not made transparent except for a brief mention of a precision value without context in the abstract. In addition, the authors have not sufficiently presented their training and test data. What they exactly mean by images that were "labelled according to their types" as training data remains unclear. While they mention the recognition of different types of images in the abstract, the actual paper is all about recognizing facial expressions. Opportunities for improvement are also not mentioned, although it is hard to imagine that they do not exist. It seems beneficial to present the experiment again on a larger scale and to make it more transparent. The information density of this paper is so low that it is hardly possible to criticize the content itself.

Pavlos Musenidis

> Simon Senecal, Louis Cuel, ... (2016): Continuous body emotion recognition system during theater performances, Computer Animation and Virtual Worlds. `https://www.researchgate.net/publication/303036265`

Motivation

The authors aim at developing a system which recognizes emotions off of body movements and maps the results onto the Russel bi-dimensional model of emotions. They want to do this in a continuous manner to also be able to show the development of the emotion in a recording over time as a trajectory on the Russel's model.

Data

To train and evaluate the system the authors chose theater performances as they are a good compromise between natural and clear expression, which is closely related to emotion. They chose eight emotions that are distributed along the four quadrants of Russel's model: *happy, excited, afraid, annoyed, sad, bored, tired* and *relaxed*. They used a Kinect for Windows v2 to capture the motion of ten different actors who each performed eight emotional states with one state ranging from 40 to 50 seconds. In these performances the actors improvise movements for the emotional state without repeating any. In total they ended up with 80 performances and approximately 53 minutes (96,000 frames) and split the data into a training set of 70%, a validation set of 15% and a test set of 15%. Additionally two actors were asked to perform a series of four specific emotions 10 seconds each to test the continuous recognition of the model.

Method

For classification they use a pipeline which consists of the Laban Movement Analysis (LMA) system, a neural network and Russel's emotion space. The LMA is a body movement description system and is used in their experiments to extract features out of the motion-captured material. They used a 35-frame sliding window with a 1-frame step to extract the LMA features which consist of measurements of body part distances, velocities, accelerations and many other body descriptions. Each feature is decomposed to four measurements: maximum (f_{max}), minimum (f_{min}), standard deviation (f_{σ}), and average (f_{μ}). This leads to a total of 87 features.The neural network consists of 86 inputs (one feature is dropped), 10 hidden layers and two outputs (coordinates for Russel's model).

Main Result

The authors do not provide a percentage of success as it is difficult because of the models continuity aspect. However their visualisation shows that the model achieves a high accuracy in placing the performed emotion in the right area in normal recognition as well as in continuous recognition. They show that LMA and neural networks are suitable for continuous analysis of emotion and estimation in terms of intensity and valence.

Critical Reflection, Limitations

The paper's main limitation is the absence of a quantitative measure for accuracy which makes it hard for other authors to compare their results in a clear quantitative manner. As the authors mention extending this system to also consider facial expressions could be a possible improvement.

Milena Voskanyan

Sapinski, T., Kaminska, D., Pelikant, A., Anbarjafari, Gh. (2019): Emotion recognition from skeletal movements, MDPI. `https://www.mdpi.com/1099-4300/21/7/646`

Motivation

In this paper, a method is proposed to recognize seven basic emotional states (happiness, sadness, surprise, fear, anger, disgust and neutral), utilizing body movement.

Data

Motion capture data used for the purpose of this research is a subset of the multi-modal database of emotional speech, video and gestures. Each recorded person was a professional actor/actress. A total of 16 people were recorded (separately). They were asked to perform the emotional states in the following order: neutral, sadness, surprise, fear, disgust, anger and happiness. All emotions were acted out 5 times.The total number of gathered samples amounted to 560 (80 samples per emotional state). Recordings took place in a quiet environment with no lighting issue. Cloud point and skeletal data feeds were captured using a Kinect v2 sensor. Data acquired from the Kinect v2 determines the 3D position and orientation of 25 individual joints. The position of each joint is defined by the vector [x, y, z], where the basic unit is 1m and the origin of the coordinate system is Kinect v2 sensor itself. The orientation is also determined with three values expressed in degrees. The device does not return orientation values of head, hands, knees and feet.

Method

From raw Kinect v2 data output, a vector containing the positions and orientation of all joints in relation to the main one was obtained. Here, for the purpose of key frame extraction, curve simplification(CS) method was used. It is assumed that every human is built in proportion to his or her height and the length of legs and arms is proportional to the overall body structure. To unify the value of the position of the joints between the higher and lower individuals, normalisation based on the distance between two joints with the lowest noise value of their position on all recordings (SpineBase and SpineShoulder) is proposed. During data preparation, a relative average quantity of motion was measured for each emotional state. The final step is classification, which aims to assign input data to a specific category k. In terms of motion emotion recognition efficiency, Convolutional Neural Network (CNN), a Recurrent Neural Network (RNN) and a Recurrent Neural Network with Long Short-Term Memory Network (RNN-LSTM) with low level features were applied in this work.

Main Result

It is observed that the best results (69 per cent) were obtained using RNN-LSTM on a set containing position of all skeletal joins (upper and lower body). In general, this set of features gives the best results for all types of networks (58.1 per cent for CNN, 59.4 per cent for RNN). This suggests that this kind of features provide the best description for emotional expressions from all considered feature types.

Critical Reflection, Limitations

In my opinion, the only limitation of the work is that only body gestures are taken into consideration here. Whereas, humans, when expressing emotions, use also facial expressions and, in some cases, even sounds. I think, in order to detect correct emotion it is necessary to take those three factors into consideration.

Wei Zhou

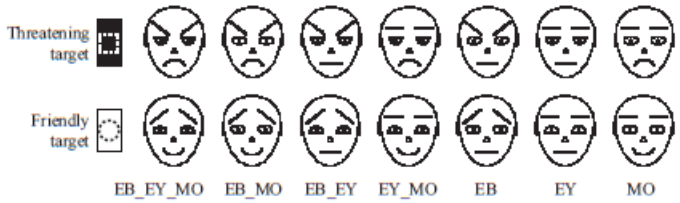
Daniel Lundqvist, Arne Ohman (2005) Emotion regulates attention: The relation between facial configurations, facial emotion, and visual attention, Visual Cognition, 12:1,51-84. https://doi.org/10.1080/13506280444000085

Motivation

1. To test whether humans preferentially orient attention towards threat.
2. To examine the relation between facial features, emotional impressions and visual attention.

Data

There are three features in the facial stimuli used in this study: the eyebrow, the eye and the mouth. Each of them has three different versions. These three versions correspond to threat, friendliness and neutral. The following figure illustrates part of the facial stimuli configurations. The figurines at row 1, column 1, row 2, column 1 and row 1, column 4 give different versions of the eyebrow. Likewise, one can find three versions of the eye and the mouth in the following figure.



Method

There are four experiments mainly differs in stimuli settings (distinguishing threatening and friendly facial stimuli with 1, 2 or 3 features. For instance, the first column in the above figure shows the setting of three different features. Within each experiment, two tasks were performed. Task 1 requires participants to decide whether there is different facial stimuli within a matrix of 9 stimuli (3*3). For instance, the target facial stimuli is the first figurine in the first row, with the rest (distracting stimuli) being the first figurine in the second row. Accuracy and reaction time were measured. In the second task, participants are asked to rate the emotion expression of facial stimuli based on activity, negative valance and potency. The independent variables are numbers of differing features, emotion expressions and type of distracting stimuli. The dependent variables are hit accuracy, reaction time and the rating scores of facial stimuli.

Main Result

1. In large cases, threatening stimuli with three, two or one features capture attention more efficiently than the corresponding friendly configuration.
2. Facial features affect both attention and emotion measures in a hierarchical way (eyebrows >large mouth >eyes). This means figures with higher ranked features are easier to get attention.
3. Strong emotional impression (high scores on activity, negative valence and potency) is associated with efficient detection (short reaction time and high accuracy).

Critical Reflection, Limitations

Schematic feature setting for threatening, friendly and neutral is too general and simple, therefore hard to be extended in real life applications.

Faizan E Mustafa

S. Saha, S. Datta, A. Konar and R. Janarthanan (2014): A Study on Emotion Recognition from Body Gestures Using Kinect Sensor , International Conference on Communication and Signal Processing. https://ieeexplore.ieee.org/document/6949798

Motivation

The goal of the study is to classify basic emotions (Anger, Fear, Happiness, Sadness and Relaxation) using the features obtained from body gestures with the help of kinetic sensors. Additionally, compare different machine learning models based on the average accuracy and computation time. The improvement in emotion recognition is expected to improve human-computer interaction.

Data

Ten subjects in the age group of 25 ± 5 were instructed to make gestures corresponding to an emotion. The total duration of 60 seconds of each emotion was acquired from each subject. The kinetic sensor acquires the data at the rate of 30 frames per second where each frame is a 3-D human skeleton represented by 20 body joints.

Method

Following nine features were extracted using eleven joints in the upper body.

- The euclidean distance between hand and elbow. (2 features)
- Maximum acceleration of hand and elbow with respect to spine. (4 features)
- Angle between head, shoulder center and spine. (1 feature)
- Angle between shoulder, elbow and wrist. (2 features)

Five models were trained using the above mentioned nine features to classify emotions. The models used were binary decision tree, AdaBoost, k-nearest neighbour, support vector machine with radial basis function kernel and neural network.

Main Result

The classification average accuracy obtained for binary decision tree, AdaBoost, k-NN, SVM and neural network classifier are 76.63%, 90.83%, 86.77%, 87.74% and 89.26% respectively. The results show that AdaBoost achieves the highest accuracy of 90.83%. However, it is worst in terms of computation time. The high average accuracy obtained for different classifiers indicate that the extracted nine features are able to distinguish between different emotions.

Critical Reflection, Limitations

The authors intend to extend the work for more difficult emotion. In my opinion, the study of body gestures is a complex topic as it depends on many factors such as culture and gender. The study does not take into account these factors and the dataset size(10 participants) is also very small. Moreover, the features were only extracted from the upper body parts. The features from lower body could also be helpful.

Yannic Becker

Vladimir J. Konečni, (2015): Emotion in Painting and Art Installations,The American Journal of Psychology , Vol. 128, No. 3 (Fall 2015), pp. 305-322. https://doi.org/10.5406/amerjpsyc.128.3.0305

Motivation

He reexamines the status of emotion in the domain of paintings.

Data

The data is a collection of different studies and theories (empirical and analytic). The data of studies is collected with probands with different group sizes from different researchers with different research questions. The capability of paintings of inducing genuine psychobiological emotions in viewers is also part of the research question.

Method

Konečni refers to his own research and summarizes and compares the studies made by other scientists. He also weaves psychological and psychobiological theories into his paper. He rates the results based on his own research and theories.

Main Result

Neither colours nor the direction of brush strokes affect the emotion. This can be seen by evaluating the reactions to the paintings of Kandinsky, Kooning and Pollock. The limited space on a canvas forces the artist to often use objects with associationist potential. Facial expressions and gestures can be in the painting, displayed by painted living objects. So the event in the painting can trigger emotions but the painting does not. The life story of the painter may also lead to emotions, when the viewer knows his story. A combination of paintings and music can increase the emotions created by the music. The knowledge around the painting, the painter and the story in the painting and around the painting make the emotions, not the canvas with the paint stuck on it. There is also the possibility that the perspective and the age of the viewer can influence the perceived emotions of the viewer by observing a painting. The author comes to the conclusion that paintings are "poor candidates for eliciting genuine psychobiological emotions" (page 317) because they have a lack of necessary details with which the viewer combines associations.

Critical Reflection, Limitations

Konečni gives a good summary of the actual (2015) research while some points stay unanswered. A lot of the studies measure the skin conductance and heart rate while none of them make the attempt to scan the cortex or interpret facial expressions and gestures. He avoids mentioning the spoken emotions of the viewer and focuses on measured data. In my opinion the measurement is one way to get data but thinking of the painting and the fact of being interested in art has a big influence on the result too. The studies often have small groups of probands and are single test constructions, which doesn't make them representative.

Marius Maile

V. Yanulevskaya, J.C. van Gemert, K. Roth, A.K. Herbold, N. Sebe, J.M. Geusebroek (2008): Emotional valence categorization using holistic image features, Proceedings of the International Conference on Image Processing. https://www.researchgate.net/publication/221122984_Emotional_valence_categorization_using_holistic_image_features

Motivation

A main intention of an artist is to capture the scene or subject such that the final masterpiece will evoke a strong emotional response. Because of this, the authors want to examine if a trained machine can perceive the main emotion evoked by these paintings. Furthermore, the perception of emotions as evoked by visual scenes is an almost untouched area of research.

Data

The training set is the International Affective Picture System (IAPS) dataset extended with subject annotations to obtain ground truth images. IAPS is a common stimulus set frequently used in emotion research. It consists of 716 categorized (positive, negative, no emotion) natural colored pictures taken by professional photographers from which 396 pictures were categorized in anger, awe, disgust, fear, sadness, excitement, contentment and amusement (ground truth images). Furthermore, a single image can belong to different emotional categories. The images portray complex scenes containing objects, people, and landscapes.

Method

They assign words to every region in the ground truth images. Because there is no vocabulary given, they assign a similarity score to all words for each region in order to leave room for uncertainty. In addition, they examine color and texture. These extracted features for each ground truth image are used to train a classifier (Support Vector Machine) to distinguish the various emotions. As the IAPS dataset is relatively small, they repeat the training and testing 25 times. Afterwards the trained system is applied to a set of masterpieces from the Rijksmuseum Amsterdam.

Main Result

The authors have shown initial results for a scene categorization system aiming to distinguish between emotional categories. Although the results are preliminary, they demonstrate the potential of machines to detect emotions evoked by master paintings by combining annotation with visual characteristics.

Critical Reflection, Limitations

The article gives a good overview of a computer-aided method for detecting emotions in images. The small data set, as already mentioned by the authors, is considered as a disadvantage. In addition, there is no visualization of how the images were decomposed by annotations and how the color and texture information was combined with them. A visualization of these steps would have been a great help. Furthermore, I think that the approach to combine annotation with image information is an innovative and goal-oriented idea, since emotion analysis as we have come to know it should never be viewed from one side only. Because of the variance of the training dataset and the chosen features, it appears to be a method that is independent of the dataset being examined.

Christina Hitzl

> Gao, Hua and Yüce, Anil and Thiran, Jean-Philippe (2014): Detecting emotional stress from facial expressions for driving safety, IEEE International Conference on Image Processing (ICIP), 5961-5965. doi:10.1109/ICIP.2014.7026203

Motivation

The motivation of this article is to contribute to more safety in road traffic by developing a method that ensures more pleasantness for the car driver. Therefore, the idea is to generate a non-intrusive method which analyzes the individual facial expressions and accordingly detect the driver's emotional state. Specifically, the facial expression the authors were looking for was stress.

Data

For the training process of the classifiers two facial expression databases (FACES and Radbound) which consist of images displaying the frontal view of faces were consulted. The data used for their evaluation is recorded with a near-infrared camera in two different settings: Firstly, the recording takes place in an office. The participants in this experiment should adapt their facial expressions such that there exist data from all six basic emotions of Ekman's model and a further neutral pose. These recordings would be useful for later model adaptations. The recordings of the second set simulate real circumstances in the car. For this scenario, fewer subjects participated.

Method

According to video recordings, the face is detected and tracked in real time and characteristics of facial expression features are collected of individual subjects. The supervised descent method is used as face tracker. Therefore, an initial shape is needed. This is done by the Viola and Jones face detector. After that, regression models are used to assign a face to the shape.
For the feature extraction they use the holistic affine warping method which normalize the face thanks to facial particularities. Secondly, local texture featues are described according to a corresponding fixed scale. Furthermore, the authors apply a pose correction method with the least squares method to avoid imbalance between training and testing data.
Finally, the approach of stress detection is made in terms of the six basic emotions of Ekman and the neutral option. Accordingly, stress is defined as anger or disgust or as a combination of both. Multi-class classifiers which are implemented with support vector machines are trained with the use of the extraced features.

Main Result

The evalution was performed on different feature extraction procedures by several methods. Each method was again tested after a model adaptation. Generally, due the pose normalization step the local descriptor approach performs better than the holistic affine warping process. Consequently, the local descriptor approach identifies 90.5 % of all videos in the first set as correct. In the second set, the detection rate of the local descriptor reaches a percentage of 85%.

Critical Reflection, Limitations

To conclude only from a facial expression to stress as the emotional state is a slightly optimistic point of view. Furthermore, to derive stress from the emotions anger and disgust may also be seen as a critical approach. Since the second experimental setting is just simulative, the testing of the model under real conditions would be more convincing.

Chat/Generation

Shawon Ashraf

> Yao, Zhang et al. (2020): Non-deterministic and emotional chatting machine: learning emotional conversation generation using conditional variational autoencoders, Neural Computing and Applications. https://link.springer.com/article/10.1007/s00521-020-05338-z

Motivation

Conversational systems like chat bots generate responses based on the text provided as input. While they can generate responses based on a specific emotion with related context, these responses are often generic and repetitive. Previous works on this topic or only based on understanding context and associated emotion and ignore the generic nature of the generated response. In this paper, the authors propose a system which tackles this problem by combining the previous works of understanding emotion and context and combining it with a non-deterministic response generator.

Data

Emotional Conversation Generation Challenge Dataset, which consists of postings and follow-up comments from the Chinese social media platform Weibo. This dataset contains more than 1 million utterance-response pairs. The datset also contains 6 emotion categories - Anger, Disgust, Happiness, Like, Sadness and Other.

Method

The authors present two models for their proposed system, NECM or Non-deterministic and emotional chatting machine. The first model, NECM-Z & E, is a concatenation of outputs from two encoders - Post encoder and Response encoder. The Post encoder approximates a representation given input text, emotion and a suitable response. The Response encoder approximates the representation from input text alone, learning the semantic features.

The second model NECM-GMM approxiates representations with Gaussian Mixture Model for all emotion categories. and generates responses using an attention based GRU or gated recurrent unit using the outputs from either NECM-Z & E or NECM-GMM as input. Also to predict the desired emotion category of the representation from NECM-GMM the authors used a Multi Layer Perceptron.

Main Result

The authors achieved lower perplexity and higher accuracy compared to existing sequence based models on automatic evaluation of generated reponses. For manual evaluation, the human evaluators had moderate agreement. Also, NECM-GMM performs better than NECM-Z & E in terms of syntactic and affective diversity, which indicates that the desired emotion for a response can be classified from input representation without supervision.

Critical Reflection, Limitations

The authors based their work on two cases of conditional probabilities. One is the probability of a representation vector given input text, a suitable response and a desired emotion category. The another is that the probability given just the input text. While this works as expected for the dataset the authors used, which is a considerable large corpus, this method may not produce expected results for smaller sized corpora or low resource languages. Also the authors do not report the metrics for MLP which classifies emotion from input.

Ching-Yi Chen

> Sayan Ghosh , Mathieu Chollet , Eugene Laksana , Louis-Philippe Morency and Stefan Scherer(2017): Affect-LM: A Neural Language Model for Customizable Affective Text Generation, 55th Annual Meeting of ACLhttps://www.aclweb.org/anthology/P17-1059.pdf

Motivation

Inspired by advances in neural language modeling and affective analysis of the text, this paper proposed a model for the representation and generation of emotional text. There are 3 research questions in the paper:

1. Can Affect-LM be used to generate affective sentences for a target emotional with varying degrees of affect strength through customizable parameter?
2. Are theses generated sentences rated as emotionally and grammatically correct ?
3. Does automatic inference of affect category from context words improve language modeling?

Data

Authors chose 4 speech corpora Fisher English Speech Corpus, Distress Assessment Interview Corpus(DAIC), SEMAINE dataset, and Multimodal Opinion-level Sentiment Intensity Dataset(CMU-MOSI) as training, fine-tuning, and evaluation data.

Method

There is an assistance model, descriptors for Affect information, which has 5 features with each feature denoting presence or absence of specific emotion. Affect-LM generates sentence consider emotion **categories** by either inferring from context using LIWC or using descriptors for affect information. Additionally, it generates sentence consider **strength** by changing model parameter to control the degree of emotion strength. There is a baseline LSTM language model. Note that authors trained and validated both Affect-LM and baseline model on Fisher dataset, and fine-tuned both model on remaining corpora. Finally, there are 2 evaluations. Firstly, estimating grammatical correctness generated sentence by crowd-sourced perception on Amazon's Mechanical Turk platform. Secondly, measured the perplexity score obtained by baseline and Affect-LM.

Main Result

The proposed Affect-LM generates texts at varying degrees of emotion strength without affecting grammatical correctness. Besides, Affect-LM is capable of generating emotionally meaningful embedding. Moreover, Affect-LM also proves that additional affective information in conversation text do improve language model prediction.

Critical Reflection, Limitations

This paper was published in 2017, so the methodology is relatively old-school, and the evaluation setting is quite simple, probably because there is no benchmark or state-of-art model in 2017. Also, it would be more realistic if Affect-LM can be applied for real dialogue generation and evaluate its performance in dialogue. However, this paper began with a gentle introduction, easy-followed approach, and then well-explained result. Readers who are new in conditional language generation like me will be able to understand the paper well.