# An Inter-Agreement Study of Automatic Emotion Recognition Techniques from Different Inputs

**Vlad-George Prundus**
Department of Computer Science
Babeş-Bolyai University
Cluj-Napoca, Romania
vlad.prundus@stud.ubbcluj.ro

**Grigoreta-Sofia Cojocar**
Department of Computer Science
Babeş-Bolyai University
Cluj-Napoca, Romania
grigoreta.cojocar@ubbcluj.ro

**Adriana-Mihaela Guran**
Department of Computer Science
Babeş-Bolyai University
Cluj-Napoca, Romania
adriana.guran@ubbcluj.ro

**ABSTRACT**
Emotions play a significant role in people's lives and interactions. However, automatic recognition of human emotions using computer systems is still a challenging task. Many approaches for automatic emotion recognition have been proposed in the last decades and a vast majority of them use only one type of input for identification, i.e. image, text, or audio. This can lead to false results as people can easily hide their emotions. In this paper, we present a study on the correlation (or inter-agreement) of the results obtained by six existing approaches for emotion recognition that process different kinds of inputs. The obtained results show that there is a low agreement between the approaches, even when they use the same type of input, and that more research is needed to determine the possible causes and also to help improve the quality of the existing tools for emotion detection.

**Author Keywords**
automatic emotion recognition, inter-agreement analysis, input sources

**CCS Concepts**

•**Computing methodologies** → **Computer vision problems;**

DOI:10.37789/rochi.2023.1.1.11

## 1. INTRODUCTION

Humans' emotions have a significant influence on their ability to perceive and comprehend the world around them. However, enabling computer systems to identify users' emotional states is a challenging task, as emotions can be manifested in different ways such as speech, gestures, and body posture, and physiological changes like blood pressure or eye gazes, and they are dependent on age, gender, or appearance. Over the last decades, a variety of approaches that try to automatically identify human emotions have been proposed. Even though there are many proposals for automatic emotion recognition with different levels of accuracy, as far as we know, there are no studies that compare the agreement between the existing approaches.

The main objective of our research is to analyze how well the existing approaches for automatic emotion recognition correlate, i.e. for a given input (e.g. a video), do all approaches obtain the same result? For this study we have selected six approaches that use different types of input (i.e., image, text, and audio), applied them to a suitable dataset, and compared the obtained results using an inter-agreement measure.

The rest of the paper is structured as follows. Section 2 introduces the domain of automatic emotion recognition and the existing approaches, Section 3 describes our study and the obtained results. Threats to validity are given in Section 4. Some remarks about related work are given in Section 5, while Section 6 concludes the paper.

## 2. AUTOMATIC EMOTION RECOGNITION

Emotion recognition is a branch of artificial intelligence that tries to automatically identify human emotions. This field is becoming more and more appealing and new improvements are constantly added to the existing approaches. Each approach tries to identify only a small set of human emotions, and, usually, the set is different from one approach to another. Most approaches identify the following emotions: anger, joy, surprise, fear, sadness, and anxiety (or a subset of these) [40].

Considering the type of input used for automatic emotion detection (or recognition), the existing approaches can be split into the following categories:

1. *Emotion detection in images*. It refers to approaches capable of identifying emotions in images or videos.
2. *Emotion detection in text*. It refers to approaches capable of identifying emotions in written text.
3. *Emotion detection in audio*. It refers to approaches capable of identifying emotions in audio recordings.
4. *Hybrid emotion detection*. It refers to approaches capable of identifying emotions from at least two different inputs.

All these categories have their own list of identified emotions and used approaches. For example, image-based emotion detection uses image-processing techniques, while text-based emotion detection uses natural language processing.

In the following, we briefly describe the first three categories, as approaches from these categories were used in our study.

### 2.1 Emotion Detection in Images

One of the most popular techniques for emotion recognition is based on visual sensors. It benefits from minimal cost and straightforward data collection. Currently, visual sensors are primarily utilized for facial expression recognition (FER) to detect emotion. The way most of the algorithms work is by taking a picture of a person, pre-processing it (to improve the quality of the image), identifying the face of the subject (or subjects), and, then, by using emotion classification, returning the overall emotion. However, these techniques suffer a major drop in accuracy, as the light intensity drops [32].

Although people's emotions can be inferred from their facial expressions, it is challenging for machines to capture the finer aspects of human expressions. Facial expressions are simple to conceal, which causes mistakes in emotion detection. For instance, even though not in a formal setting, people often pleasantly smile during several social events [21]. Also, people have different skin tones, appearances, and facial features, making difficult any attempts to accurately classify them. The facial expressions of the same emotion might vary, and subtle differences across the emotions of the same person are not always visible. When the face is obscured (such as when wearing a mask) or when the camera is angled differently, it is also challenging to accurately identify emotions [11].

### 2.2 Emotion Detection in Text

These days, text-based devices are widely and efficiently utilized for communication. Also, as a result of the rapid rise of social media, users frequently express their emotions, opinions, and feelings on social media sites like Twitter, Facebook, or YouTube. Although many social media users convey their feelings using audio and video, written text is still the preferred way. Through postings, status updates, comments, and blogs on social media, people frequently convey their feelings. To determine what emotions are being expressed in these posts, an analysis of them is required. Understanding emotional cues is essential for social interactions because they help us understand others' mental states and behavioral responses [33].

There are three different types of emotion detection methods from written text: deep learning, machine learning, and lexicon-based methods, each having advantages and disadvantages. The general approach used for emotion detection from written text usually contains the following stages: collecting the dataset (used for training), pre-processing (tokenization, normalization, removing stopwords, stemming, lemmatization), feature extraction (bag of words, Ngram, word embedding), model development (using deep learning or machine learning approaches), and model assessment. However, despite various methods for recognizing emotions from written text, it is still a challenge to deal with context, mockery, statements that express multiple emotions, the growth of Web slang, and lexical and syntactic ambiguity. Furthermore, because there are no established guidelines for conveying emotions across many media, some people express their feelings to amazing effect, while others repress them. Therefore, creating a method that is effective across all domains is still a big challenge.

### 2.3 Emotion Detection in Audio

One of the crucial aspects of human culture is spoken language. Language is a means of communication by which people can express themselves or interact with others. Speech emotion recognition (SER) has also advanced thanks to voice recognition. For the purpose of identifying emotions, human speech offers a wealth of information. For artificial intelligence-based systems to communicate effectively with humans, it is crucial to comprehend the emotions contained in the information. SER can be used for autism diagnosis, automatic response systems, call-center dialogue, and more. Acoustics feature extraction and language mark work together to accomplish SER[3].

For automatic emotion recognition from speech, during the pre-processing stage, the following steps are carried out: noise reduction and enhancement of the input signal into segmentation, feature extraction, and classification. Afterward, with certain semantic contributions, the language model may recognize emotional expressions. By examining prosodic or spectral characteristics, the acoustic model is able to identify several emotions that are present in a single utterance [8].

However, understanding emotions in speech is a complex process. Different speaking styles of various people result in acoustic variability, which directly affects speech feature labeling and extraction. The same sentence may contain different emotions, and some specific emotional differences often depend on the speaker's local culture or living environment, which also pose challenges for SER [20].

## 3. STUDY

The purpose of our study is to provide an inter-agreement analysis of some of the existing approaches for emotion detection using three of the most commonly used inputs: images (facial expression), audio (speech audio in English), and text (speech text in English). For the analysis to be relevant and provide useful information we have decided to use two different approaches for each type of input channel:

1. For *image* we have used Facial expression recognition (FER) [35] and DeepFace [34] libraries.

2. For *audio* we have used a trained Multi-layer Perceptron classifier (MLPClassifier) [29] and a trained classifier based on a Sequential model [24].

3. For *text* we have used ParallelDots API [28] and the text2emotion library [1].

In the following, we briefly describe the selected approaches.

*Facial expression recognition (FER)* [35] is a free Python library used for identifying emotions either in an image or in a video. The identification of emotions in the video is done by splitting the recording into frames and identifying the emotions in each image. It identifies the face area in the image by using OpenCV's Haar Cascade classifier (or

even a Multi-Task Cascaded Convolutional Neural Network if needed). The model used for emotion detection is a convolutional neural network with weights saved in an HDF5 file. The accuracy of the model is around 50% [5].

*DeepFace*[34] is a free Python library used for facial analysis in images. It can be used for emotion detection, face recognition, and facial attribute analysis (gender, race, age). Similar to FER, it uses OpenCV for face detection, but it does not have a function for splitting a video into frames, so the user must manually send each image from the recording. However, DeepFace can provide real-time analysis for a live recording and display the attributes on the screen (emotions, gender, age). For facial attribute analysis DeepFace uses convolutional neural network (CNN) models with an accuracy of 80% [2, 19, 38, 31].

The *MLPClassifier* [29] we have used for audio analysis is a pre-trained model that is able to identify the emotions of the speaker using the audio recording of their speech. The classifier was trained using the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset [23] which contains 24 professional actors, half male, half female vocalizing two statements. The dataset contains 7356 files, that are available either audio only or with video. The files are also available either as a song or as a speech. The spoken language is English with a North American accent. Each file contains an emotion from the following: calm, happy, sad, anger, fearful, surprise, and disgust together with the intensity level (normal, strong). The labels were given by 247 individuals from North America [37]. For the MLPClassifier we have considered three features: *mfcc* - Mel Frequency Cepstral Coefficient (represents the short-term power spectrum of a sound), *chroma* - pertains to the 12 different pitch classes, and *mel* - Mel Spectrogram Frequency. The accuracy of the model is 60.26% [27]. This approach is denoted as Trained Audio in our study.

The *trained classifier based on a Sequential model* is a pre-trained model used for the identification of emotions in audio. It has been trained using three datasets: RAVDESS [23], Toronto emotional speech set (TESS) [25] and Surrey Audio-Visual Expressed Emotion (SAVEE) [24]. The TESS dataset contains two actresses pronouncing the sentence "Say the word ..." together with 200 words. Each recording was done by portraying one of the following emotions: anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral. The spoken language is English and both speakers have the language as their native one [22]. The SAVEE dataset contains four native English speakers aged from 27 to 31 years telling a total of 480 sentences in six emotions: anger, disgust, fear, happiness, sadness, and surprise [30]. This model uses a series of speech features such as frequency, amplitude, and decibels to map the emotion to the audio. The overall accuracy of the model is 84.96% [14]. This approach is denoted as Speech Analyzer in our study.

*ParallelDots API* [28] is a paid API that can identify emotions in text. The ParallelDots company delivers a series of APIs and solutions for lots of needs in text analysis such as sentiment analysis, semantic analysis, emotion detection, text classification, intent analysis, and many more. Although their accuracy is not public, they state that the software uses a deep learning model on their own dataset, providing high accuracy, fast analysis, and flexible deployment. They also support multiple languages besides English such as: Spanish, Italian, German, French, Japanese, and so on. Although it is a paid API, it features a free trial for 30 days with 1000 requests per day [13].

*Text2Emotion*[1] is a free Python library that uses a lexicon-based approach to identify emotions in text, so it does not need training as the previous approach and can be used directly on any English text. It uses a dictionary where every word contains an emotion, and the result of the algorithm is the emotion most present in the text. There are no public records with the accuracy of the library but from previous experiences, we found out to be around 70%. For testing the accuracy, we have used a dataset that contains 1800 texts with emotions such as fear, joy, anger, sadness, and love [6].

### 3.1 Identified Emotions

For our inter-agreement analysis to be accurate we needed to have the same output for each approach. As there are so many possible emotions that can be identified, we have decided to use the most common ones: *anger*, *happy*, *disgust*, *fear*, *sad*, *surprise*, and *neutral*. Most of the selected approaches already return these emotions, so we had to update only one of them, namely the *text2emotion* approach. The *text2emotion* approach returns only five emotions: happy, anger, sad, surprise, and fear. It is missing disgust and neutral. For neutral emotion, we have decided to return it when the model is returning the dictionary of emotions with every feeling on the 0 value (there are no emotions available, so it is a neutral emotion). For disgust emotion, we have used the NRC Word-Emotion Association Lexicon (EmoLex) [39]. The EmoLex is a dictionary with over 14.000 words from 108 languages, divided into two sentiments (negative and positive) and eight emotions (anger, anticipation, disgust, fear, joy, sadness, surprise, and trust). To complete the emotions for *text2emotion* we took all the words containing the disgust feeling from the lexicon and added them to the existing ones in the model.

### 3.2 Dataset

In order to be able to compare the results, we must use the same dataset as input for each selected approach. However, there are very few public or free datasets available for the purpose of our analysis. To be able to compare all three inputs we need a dataset that contains videos with lengths between one and five minutes, with only one person that is speaking. These restrictions are imposed because of the nature of each approach. In the case of emotion detection in videos, we need to extract and analyze every frame, identify the emotions and then return an overall emotion. This proves to be very time-consuming since a video of three minutes, filmed with 60 frames/second will result in over 10.000 images that need to be analyzed one by one. As such, we have restricted the video length to a maximum of five minutes. The minimum video length of one minute is due to the emotion analysis in audio and text. In order to be conclusive, the speech should be longer than just a few words. So, we need videos that have

at least one minute in length. Also, the videos should contain only one subject since it is easier to identify the emotions of only one person. For the purpose of this study, we did not choose to also compare the emotions of two people at the same time. Lastly, we need to hear only one person so an interview, even if we can see only one person, is not suited for our research. The only dataset that was free and available was the Aff-Wild2 dataset [16]. The dataset contains 564 videos with around 2.8M frames, 326 male subjects, and 228 female subjects. The videos are taken from the internet, containing videos with real people (not actors), dealing with many different situations like playing games, watching TV, having a pitch, vlogging, etc. All 564 videos have been annotated in terms of valence and arousal, while 546 of around 2.6M frames have been annotated in terms of emotions. The emotions used for annotations are the ones used in this research: *anger*, *happy*, *fear*, *sad*, *disgust*, *surprise*, and *neutral*.

After analyzing all the videos available in the dataset, we selected **152** videos with a total length of over five hours. The number of samples per emotion category are given in Table 1. The videos respect the restrictions imposed earlier and contain both male and female subjects with ages between 20 and 70 speaking in English.

Table 1: The number of selected videos per emotion.

| Emotion | Number of videos |
|---|---|
| angry | 16 |
| disgust | 5 |
| fear | 13 |
| happy | 73 |
| sad | 19 |
| surprise | 5 |
| neutral | 21 |
| **Total** | 152 |

### 3.3 Results

After selecting the approaches, the emotions, and the dataset the next step was to obtain the results. For this purpose, we needed to prepare the input data for each algorithm, as it follows:

- For *emotion detection in videos* we used Open Source Computer Vision Library (OpenCV) from Python, which allows to both split the video into frames and save the resulting images in a folder. Each image from that folder was sent to the algorithm, which returns the most intense emotion found (the emotion with the highest score). The overall emotion of the video was considered to be the emotion that is the most frequent in the video (the emotion that is present in almost every frame).
- For *emotion detection in audio* each video was converted from an mp4 format to a wav format. That way only the audio is sent to the algorithm which returns the dominating emotion of the whole audio. This was done using the AudioSegment [4] library in Python that takes the video file path and saves the new audio file in the given destination path.
- For *emotion detection in text* each video needed to be transcribed from audio to text. In our case we have used Deepgram [7], which is one the most powerful speech-to-text tools available, having an accuracy of over 85%. They support many audio or video formats as well as any length for the file sent. The obtained result contains various information about the request together with the transcript text, the confidence, and every word identified together with their confidence. Also, each word has a start and an end time (in seconds)[23].

The results of the experiment are presented in Table 2. For each approach and each emotion, it shows the number of videos for which the emotion was correctly identified. Although the results are useful, they only show the accuracy of the approach for the selected dataset. Still, there are some unexpected accuracy values. For example, DeepFace which should have had an accuracy close to 80% only got 28% accuracy while FER with a general accuracy of 50% got 40% in our tests. These results might have multiple causes such as the quality of the videos and audio, the number of emotions a person is experiencing (if a video is five minutes long a person might be experiencing many emotions if talking about multiple subjects), and, also, the ability of a person to hide his/her emotions.

For our study, we are interested to determine the correlation (or inter-agreement) of the obtained results, i.e. for a given video, how many of the selected approaches have identified the same emotion. As such, we have used the Fleiss Kappa measure that computes the degree of agreement in classification over that which would be expected by chance. The formula for this measure is given in Equation 1:

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \tag{1}$$

where, given $N$ cases (or subjects) which are each rated by $n$ raters, and each rater can give one value from a set of $k$ possible values, $\bar{P}$ is the observed agreement between the raters and $\bar{P}_e$ is the expected agreement if raters make random judgments [10]. The values of $\kappa$ are between 0 and 1, and a possible interpretation for the results obtained by this measure is shown in Table 3 [17]. However, this interpretation is not generally accepted, as the obtained results are dependent on the number of categories and cases.

In order to compute this measure, we have used the Online Kappa Calculator available at http://justusrandolph.net/kappa/. In our case, we have 152 cases (the total number of videos from our dataset) and 7 categories (the number of possible emotions: *anger*, *disgust*, *fear*, *happy*, *sad*, *surprise*, and *neutral*). The number of raters depends on the combination used: 2 raters when we compare the results obtained by approaches using the same type of input (Table 4), 2 raters when we compare the results obtained by two approaches using different types of input (Table 5), 3 raters when we compare the results obtained by three approaches using three different types of input (Table 6), and 6 raters when we compare the results obtained by all six approaches (Table 7).

Table 2: Correctly identified emotions for each approach

| Emotion | FER | Deep Face | Parallel Dots | Text2 Emotion | Trained Audio | Speech Analyzer |
|---|---|---|---|---|---|---|
| **anger** | 11 | 4 | 3 | 1 | 3 | 0 |
| **disgust** | 0 | 0 | 0 | 0 | 2 | 3 |
| **fear** | 0 | 4 | 2 | 3 | 4 | 0 |
| **happy** | 31 | 16 | 29 | 6 | 3 | 6 |
| **sad** | 8 | 7 | 2 | 7 | 2 | 0 |
| **surprise** | 0 | 0 | 0 | 2 | 0 | 0 |
| **neutral** | 0 | 0 | 0 | 0 | 0 | 0 |
| **Accuracy** | 40% | 28% | 26% | 16% | 11% | 6% |

Table 3: Fleiss Kappa Measure Interpretation[17]

| $\kappa$ | Interpretation |
|---|---|
| $< 0$ | No agreement |
| 0.01-0.20 | Slight agreement |
| 0.21-0.40 | Fair agreement |
| 0.41-0.60 | Moderate agreement |
| 0.61-0.80 | Substantial agreement |
| $> 0.80$ | almost perfect agreement |

Table 4: Fleiss Kappa values for the same type of input

| Pair | Measure value |
|---|---|
| FER + DeepFace | **0.4145**% |
| Parallel Dots + Text2Emotion | 0.2566% |
| Trained Audio + Speech Analyzer | 0.1184% |

Table 5: Fleiss Kappa values for two different inputs

| Combination | Measure value |
|---|---|
| FER+ParallelDots | 0.2434% |
| FER+Text2Emotion | 0.2434% |
| FER+Trained Audio | 0.0789% |
| FER+Speech Analyzer | 0.0461% |
| DeepFace+ParallelDots | **0.2961**% |
| DeepFace+Text2Emotion | 0.2763% |
| DeepFace+Trained Audio | 0.1842% |
| DeepFace+Speech Analyzer | 0.0658% |
| ParallelDots +Trained Audio | 0.1118% |
| ParallelDots +Speech Analyzer | 0.0987% |
| Text2Emotion +Speech Analyzer | 0.0724% |
| Text2Emotion +Trained Audio | 0.2303% |

As can be seen in Tables 4, 5, 6 and 7 the highest agreement obtained is *Moderate agreement* for a combination of two approaches (FER+DeepFace), with the same type of input (images). When comparing the results obtained by two or three different types of input, the highest agreement level is *Fair agreement* for emotions identified from images (using DeepFace) and text (using ParallelDots). If for combinations of two different types of inputs, 5 out of 12 combinations are having a *Fair agreement*, when we combine three different types of inputs, only one combination has *Fair agreement*, and 7 of them reached only *Slight agreement*. If we compare the results obtained by all six selected approaches, the level of agreement is only *Slight agreeement*.

From these results, we can conclude that different types of input for the emotion recognition approaches yield different results, and further studies are needed to determine the possible causes.

### 3.4 EmoFinder Tool

In order to support researchers identify the parts of a video where the selected approaches disagree we have developed an application that visualizes the results obtained by each approach on a smaller interval (e.g. 10 seconds, 20 seconds, etc). It allows the user to select an approach for each type of input and the video for the analysis, and it saves the results to a file. The file contains each identified emotion for every interval. Also, a graphic comparing the emotions detected by the selected approaches (as shown in Fig. 1) is provided.

## 4. THREATS TO VALIDITY

For our study, we have identified the following threats that could lead to invalid results:

- *Quality of videos*. The quality of videos can impact the results of the approaches. A video can have its audio low quality, while the image high quality, and vice versa. This can impact one or more inputs and result in inconsistent results, with a low agreement, but can be overcome by using approaches that perform well even in low-quality conditions. In our study, the videos used were part of only a few data sets available for emotion detection. Their raw and unedited nature made it possible to test the approaches in real life and proved useful in day-to-day situations, rather than in simulated situations.

Table 6: Fleiss Kappa values for three different inputs

| Combination | Measure value |
|---|---|
| FER+ParallelDots+Trained Audio | 0.1442% |
| FER+ParallelDots+Speech Analyzer | 0.1294% |
| FER+Text2Emotion +Trained Audio | 0.1842% |
| FER+Text2Emotion +Speech Analyzer | 0.1206% |
| DeepFace+ParallelDots+Trained Audio | 0.1974% |
| DeepFace+ParallelDots+Speech Analyzer | 0.1535% |
| DeepFace+Text2Emotion +Trained Audio | **0.2303**% |
| DeepFace+Text2Emotion +Speech Analyzer | 0.1382% |

Table 7: Fleiss Kappa values for all six approaches

| Combination | Measure value |
|---|---|
| FER+ParallelDots+Trained Audio+ DeepFace+Text2Emotion +Speech Analyzer | 0.1825% |

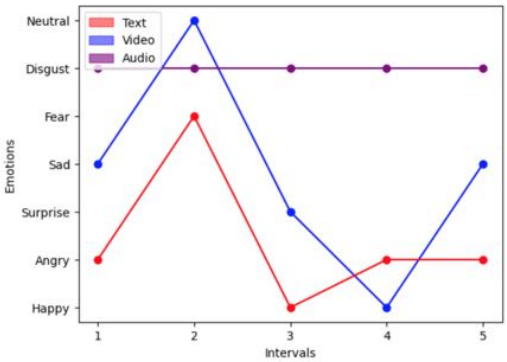


Figure 1: Graphic of identified emotions.

- *Number and selection of videos*. The selection of the videos for our study was done to facilitate the identification of emotions. However, this resulted in only approx. 150 videos. The number and subjects of videos can impact the quality of the comparison. More videos with a wider range of emotions and interactions could result in a better correlation of the approaches. Still, because there are very few data sets available online, it proves difficult to create a larger test set.

- *Number and type of chosen approaches*. The study has only focused on three types of input and two approaches for each type of input. However, these selected approaches represent only a small amount of available tools that can be used to identify emotions. Because of their free and open-source nature, the results obtained might not result in a conclusive agreement or disagreement. The choice of approaches was done due to emotions identified, price, and time constraints: we needed to use free tools or tools that have a free trial version and tools that can be easily integrated into a Python application. The approaches also required to have similar or identical emotions identified. With more resources, the number and types of used approaches can be extended.

## 5. RELATED WORK

There are currently many papers that discuss the subject of emotion detection for all three inputs: text [12] [15], audio [26], video [18] [36]. However, very few researchers have experimented with emotion detection from multiple channels, usually with text and speech, mainly because of their close nature [41]. Dupre et al. tested eight commercially available automatic classifiers and compared their emotion recognition performance to that of human observers [9]. They focused on the following emotions: anger, disgust, fear, happiness, sadness, and surprise and their results showed a recognition advantage for human observers over automatic classification, and the best recognition accuracy for an automatic classifier was 62%. Still, we did not find any research in the literature that performs a correlation or inter-agreement analysis of the results obtained by different emotion recognition approaches.

## 6. CONCLUSIONS AND FURTHER WORK

We have presented in this paper our correlation study on the results obtained by different emotion recognition approaches, using different types of input (i.e. image, audio, and text). The obtained results show that in most cases there is only slight agreement, some combinations obtain a fair agreement, and only one combination manages to reach a moderate agreement. There are many causes for these results, but further studies are needed in order to determine the exact reasons and to improve the correlation. In order to help researchers identify the causes, we have developed a tool that, for a video and a preset interval, computes the results of different emotion recognition approaches.

In the future, we intend to:

- Add more approaches for our inter-agreement analysis.
- Try other statistical measures for computing the inter-agreement.
- Add more input channels to our EmoFinder tool.

## REFERENCES


[1] S. Sharma A. Gupta, A. Band and K. Bilakhiya. 2023. Text2Emotion. https://shivamsharma26.github.io/text2emotion/. (2023). last accessed: march 2023.

[2] Asad Abdi, Siti Mariyam Shamsuddin, Shafaatunnur Hasan, and Jalil Piran. 2019. Deep learning-based sentiment classification of evaluative text based on Multi-feature fusion. *Information Processing & Management* 56, 4 (2019), 1245–1259.

[3] Jeremy Ang, Rajdip Dhillon, Ashley Krupski, Elizabeth Shriberg, and Andreas Stolcke. 2002. Prosody-based automatic detection of annoyance and frustration in human-computer dialog.. In *INTERSPEECH*, John H. L. Hansen and Bryan L. Pellom (Eds.). ISCA, 2037–2040.

[4] AudioSegment. 2023. AudioSegment's documentation. https://audiosegment.readthedocs.io/en/latest/index.html. (2023). last accessed: march 2023.

[5] Anil Bandhakavi, Nirmalie Wiratunga, Deepak Padmanabhan, and Stewart Massie. 2017. Lexicon based feature extraction for emotion text classification. *Pattern Recognition Letters* 93 (2017), 133–142. Pattern Recognition Techniques in Data Mining.

[6] Ankush Chatterjee, Umang Gupta, Manoj Kumar Chinnakotla, Radhakrishnan Srikanth, Michel Galley, and Puneet Agrawal. 2019. Understanding Emotions in Text Using Deep Learning and Big Data. *Computers in Human Behavior* 93 (2019), 309–317.

[7] Deepgram. 2023. The most powerful speech-to-text API. https://deepgram.com/. (2023). last accessed: march 2023.

[8] F. Dellaert, T. Polzin, and A. Waibel. 1996. Recognizing emotion in speech. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, Vol. 3. 1970–1973 vol.3.

[9] Damien Dupré, Eva G. Krumhuber, Dennis Küster, and Gary J. McKeown. 2020. A performance comparison of eight commercially available automatic classifiers for facial affect recognition. *PLoS ONE* 15, 4 (2020). DOI: http://dx.doi.org/https://doi.org/10.1371/journal.pone.0231968

[10] J.L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76, 5 (1971), 378–382.

[11] Deepak Ghimire and Joonwhoan Lee. 2013. Geometric Feature-Based Facial Expression Recognition in Image Sequences Using Multi-Class AdaBoost and Support Vector Machines. *Sensors* 13, 6 (2013), 7714–7734.

[12] Vidyasagar Potdar Haji Binali, Chen Wu. 2010. Computational approaches for emotion detection in text. In *Institute of Electrical and Electronics Engineers*.

[13] Maryam Hasan, Elke A. Rundensteiner, and Emmanuel O. Agu. 2018. Automatic emotion detection in text streams by analyzing Twitter data. *International Journal of Data Science and Analytics* 7 (2018), 35 – 51.

[14] Vinay Kumar Jain, Shishir Kumar, and Steven Lawrence Fernandes. 2017. Extraction of emotions from multilingual text using intelligent text processing and computational linguistics. *Journal of Computational Science* 21 (2017), 316–326.

[15] Edward Chao-Chun Kao, Chun-Chieh Liu, Ting-Hao Yang, Chang-Tai Hsieh, and Von-Wun Soo. 2009. Towards Text-based Emotion Detection A Survey and Possible Improvements. In *Institute of Electrical and Electronics Engineers*.

[16] Dimitrios Kollias and Stefanos Zafeiriou. 2018. Aff-Wild2: Extending the Aff-Wild Database for Affect Recognition. *CoRR* abs/1811.07770 (2018). http://arxiv.org/abs/1811.07770

[17] J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33, 1 (1977), 159–174.

[18] R. Nakatsu L.C. De Silva, T. Miyasato. 1997. Facial emotion recognition using multi-modal information. In *Institute of Electrical and Electronics Engineers*.

[19] Quoc Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In *Proceedings of the 31st International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Eric P. Xing and Tony Jebara (Eds.), Vol. 32. PMLR, Bejing, China, 1188–1196.

[20] Shi-wook Lee. 2019. The Generalization Effect for Multilingual Speech Emotion Recognition across Heterogeneous Languages. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 5881–5885.

[21] Shan Li and Weihong Deng. 2022. Deep Facial Expression Recognition: A Survey. *IEEE Transactions on Affective Computing* 13, 3 (2022), 1195–1215.

[22] Zongxi Li, Haoran Xie, Gary Cheng, and Qing Li. 2021. Word-level emotion distribution with two schemas for short text emotion classification. *Knowledge-Based Systems* 227 (2021), 107163.

[23] S.R. Livingstone and Russo F.A. 2018. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLoS ONE* 13, 5 (2018).

[24] EU JIN LOK. 2019a. Surrey Audio-Visual Expressed Emotion (SAVEE). https://www.kaggle.com/datasets/ejlok1/surrey-audiovisual-expressed-emotion-savee. (2019). last accessed: march 2023.

[25] EU JIN LOK. 2019b. Toronto emotional speech set (TESS). https://www.kaggle.com/datasets/ejlok1/toronto-emotional-speech-set-tess. (2019). last accessed: march 2023.

[26] Gunes Karabulut Kurt Mehmet Cenk Sezgin, Bilge Gunsel. 2012. Perceptual audio features for emotion detection. *EURASIP Journal on Audio, Speech, and Music Processing* 16 (2012).

[27] Saif M. Mohammad and Peter D. Turney. 2013. CROWDSOURCING A WORD–EMOTION ASSOCIATION LEXICON. *Computational Intelligence* 29, 3 (2013), 436–465.

[28] ParallelDots. 2023. ParallelDots AI APIs. https://apis.paralleldots.com/text_docs/index.html. (2023). last accessed: march 2023.

[29] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[30] Soujanya Poria, Alexander Gelbukh, Erik Cambria, Amir Hussain, and Guang-Bin Huang. 2014. EmoSenticSpace: A novel framework for affective common-sense reasoning. *Knowledge-Based Systems* 69 (2014), 108–123.

[31] Tapasy Rabeya, Sanjida Ferdous, Himel Suhita Ali, and Narayan Ranjan Chakraborty. 2017. A survey on emotion detection: A lexicon based backtracking approach for detecting emotion from Bengali text. In *2017 20th International Conference of Computer and Information Technology (ICCIT)*. 1–7.

[32] Recfaces. 2023. Emotion recognition: introduction to emotion reading technology. https://recfaces.com/articles/emotion-recognition. (2023). last accessed: 10 04 2023.

[33] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 815–823.

[34] S. I. Serengil. 2023. DeepFace. https://github.com/serengil/deepface. (2023). last accessed: april 2023.

[35] J. Shenk. 2023. FER. https://github.com/justinshenk/fer. (2023). last accessed: april 2023.

[36] Mohammad Soleymani, Sander Koelstra, Ioannis Patras, and Thierry Pun. 2011. Continuous emotion detection in response to music videos. In *Institute of Electrical and Electronics Engineers*.

[37] Carlo Strapparava and Alessandro Valitutti. 2004. WordNet Affect: an Affective Extension of WordNet. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. European Language Resources Association (ELRA), Lisbon, Portugal.

[38] Songbo Tan and Jin Zhang. 2008. An empirical study of sentiment analysis for chinese documents. *Expert Systems with Applications* 34, 4 (2008), 2622–2629.

[39] Guangxia Xu, Weifeng Li, and Jun Liu. 2020. A social emotion classification approach using multi-model fusion. *Future Generation Computer Systems* 102 (2020), 347–356.

[40] SoYeop Yoo, Jein Song, and Okran Jeong. 2018. Social Media Contents based Sentiment Analysis and Prediction System. *Expert Systems with Applications* 105 (2018), 102–111.

[41] Seunghyun Yoon, Seokhyun Byun, and Kyomin Jung. 2018. Multimodal Speech Emotion Recognition Using Audio and Text. In *2018 IEEE Spoken Language Technology Workshop (SLT)*. 112–118.