Goldfish of Solitude: a backchanneling chat companion for the elderly with supportive reporting for caretakers.

Alexandra Veller

Ana Daniela Peres Rebelo

# Introduction

Elderly people are often lonely because their life partners have passed away and often so have their friends who are usually age peers. They often have reduced mobility as well, and it becomes more and more difficult for them to go to places for social contact. Often, elderly people start talking to themselves or to familiar objects that they may fixate on (e.g. pictures of their beloved). Elderly people often have pets as their sole regular social companion, but taking care of a pet sometimes becomes impossible, especially when a move to a care facility is required.

Because pets seem to be a somewhat adequate substitute for human companionship, we have designed our interface with an “artificial pet” in mind. It is our impression that people, and especially elderly people do not like to be bossed around. Not by others, but certainly not by technology. With pets however, they are in control, and even if the animal is in control, they will not mind this much. Pets are usually seen as non-threatening to individuality or independence.

This type of interface is also very suitable for people who are not well-versed with technology or who are losing cognitive capacity. The elderly end-user just looks at the fish and talks to the fish, there is no typing or other manipulation required.

Because this is a student project with limited time at hand, we limit ourselves to presenting a visual and auditory (not tactile) interface only. Inspired by the song “Goudvis van Eenzaamheid” (Goldfish of Solitude) by Senne Guns, wherein an 83-year-old lady pours her heart out to her goldfish, we have created a simple interface representing a fish tank with a mobile and dirigible goldfish.

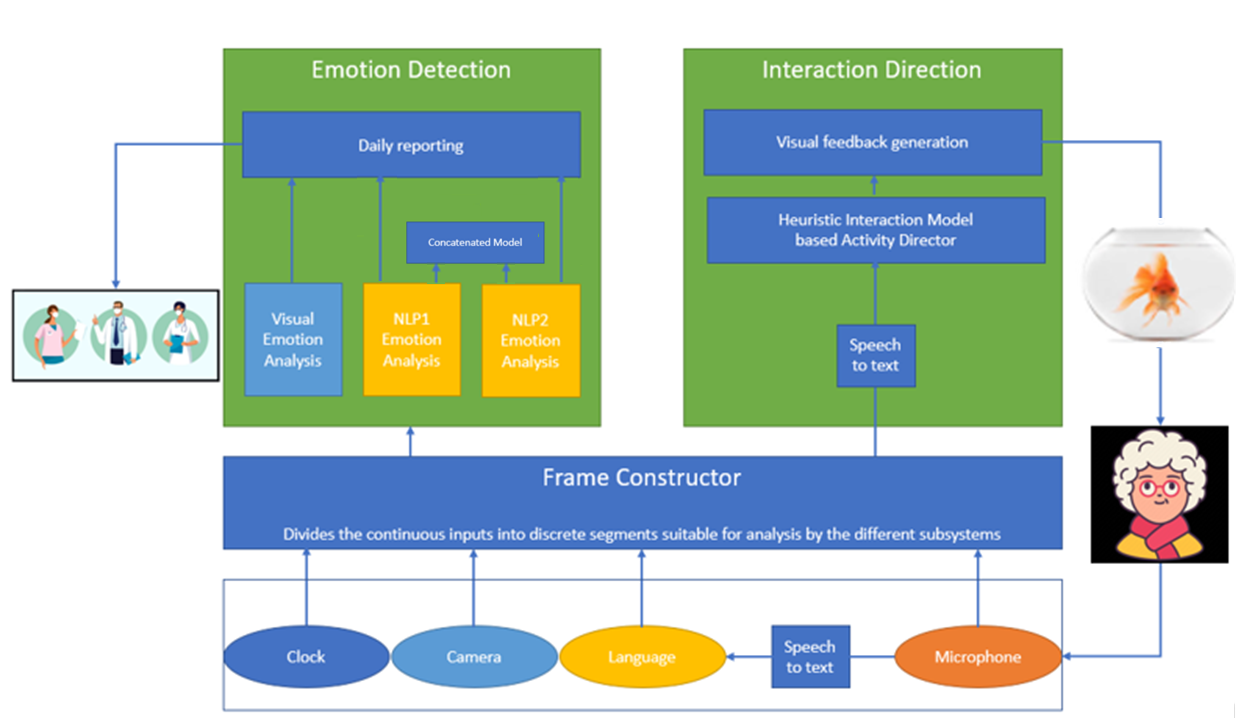
The choice for the goldfish was motivated by several factors:

1. In real life, goldfish are confined to a tank. The movements of the fish are limited to the confines of the tank. This allows us to create a credible interface, confined to a limited amount of space, which reduces the complexity of the project.
2. In real life, goldfish are not creatures people hug or pet. Their watery habitat is separate from humans, hereby making the absence of physical contact credible, again reducing the complexity of the project.
3. In real life, goldfish are not known for their communicative skills. Expectations of their abilities are therefore naturally low; they will not be expected to comment on the color of the new curtains. This again reduces the complexity of the project.
4. While real life fish have limited qualities as pets, people do keep them for their decorative value. Fish tanks with tropical fish are colorful and full of movement and life. They are both soothing and stimulating at the same time. This can be a bonus for elderly people with waning cognitive capabilities such as is the case for Alzheimer patients or patients with dementia.
5. While communication with real fish is in principle limited, our artificial goldfish will not be entirely as uncommunicative as real-life fish are. Our fish will be able to show attention when spoken to and will have a limited palette of responses that are not true-to-life. We hope that this will come across as credible, in view of the fact that our goldfish is a cartoon goldfish and cartoon animals are known to behave differently from real animals.

Our interface was created with Inkscape and Unity. Both the background image and the shape and color of the fish were chosen based on input from professionals working with the elderly, with expertise in the areas of Alzheimer and dementia. We have chosen for a colorful, stimulating background and the shape and color of the fish have been chosen for optimal likability (round shape, big eyes; mimicking baby mammal features) and color contrast with the background (maximizing visibility and resolution).

# Project Overview – NLP point of view

This project was carried out in the context of the course module Intelligent Interfaces of the postgraduate AI at Erasmushogeschool Brussel. The participants to the project were Ilja De Rycke, Reginald Van Woensel, Ana Daniela Peres Rebelo and Alexandra Veller.

These project notes reflect the NLP part of the project as carried out by Alexandra Veller and Ana Daniela Peres Rebelo.

*Figure 2 – Project Overview*

Figure 1 above shows the overview of the project, whereby inputs into the system will be threefold:

1. Input from camera (video)
2. Input from microphone
3. Input from language (NLP based on STT from transformed audio input)

The formatted inputs are then fed into two separate systems with separate goals.

*System1: Emotion Detection and Reporting*

The raw inputs are fed to three different systems: the visual analysis and the analyses based on NLP systems 1 and 2.

We decided to use two NLP models for two reasons:

1. At the onset of the project, we had selected BERT as state-of-the-art basis model for our NLP brain bit. Professor Middag however warned us that older systems might do an equally good job and would be easier to handle. We therefore implanted two different NLP systems.
2. During the course of our research, we found out that BERT pretrained models may not actually do what we think they do. It would therefore be interesting to see what the results would be based on another model, as this model may, in the end, be more reliable than a BERT-based model.

The emotion analysis results are sent to a diary that notes the emotions and the times of day at which they occurred. The diary can be accesses by the medical caretaker of the elderly person who, in the context of what is already known about the elderly person in their care, their context and history, can interpret the data and decide if action or intervention is warranted.

*System 2: Interaction Direction*

This system decides when to interact and what kind of behavior the fish should display. System 2 oversees the backchanneling.

While working on the project the setup changed somewhat to accommodate for feasibility. System 1 and System 2 are now separate from each other and the Emotion Detection system only feeds into the diary that is sent (at the end of the day and not real-time) to the caregiver/medical professional and does not feed into the backchanneling system. The reason for this adaptation is that it is impossible for us to perform the multimodal emotion detection in real time. System 1 can therefore not be used for backchanneling, which presupposes near-to-real-time interaction with the end-user.

The backchanneling relies on speech input only, which is translated to text by a (Google) speech-to-text service. The written input is then checked for key words by a lexicon-based and rule-based system. Both the lexicon and the rule-based processing are proprietary and created by the team. The Interaction Direction systems checks the textual input for emotion keywords and triggers actions for the fish interface based on the output emotion of the rule-based system.

*Multimodality – concatenation of models*

At the onset we wanted to make a multimodal model that would be able to draw conclusions from multiple input modes (video, NLP, and prosody). Due to practical issues this aim could not be pursued to the end. Prosody analysis proved to be exceedingly difficult and because the assignment only required us to use two brain bits, we opted to divert our time to the backchanneling instead, in order to have a fully functional system that we would be able to demo and that still complies with the assignment conditions.

Because we did not have the prosody information, we had to limit ourselves to the modalities of video and NLP. The NLP team investigated how to concatenate the results of two models and how to train a new model on the basis of these inputs. In the end the NLP team did not receive the output data from the Vision system, and we were unable to proceed with the multimodal part of the project.

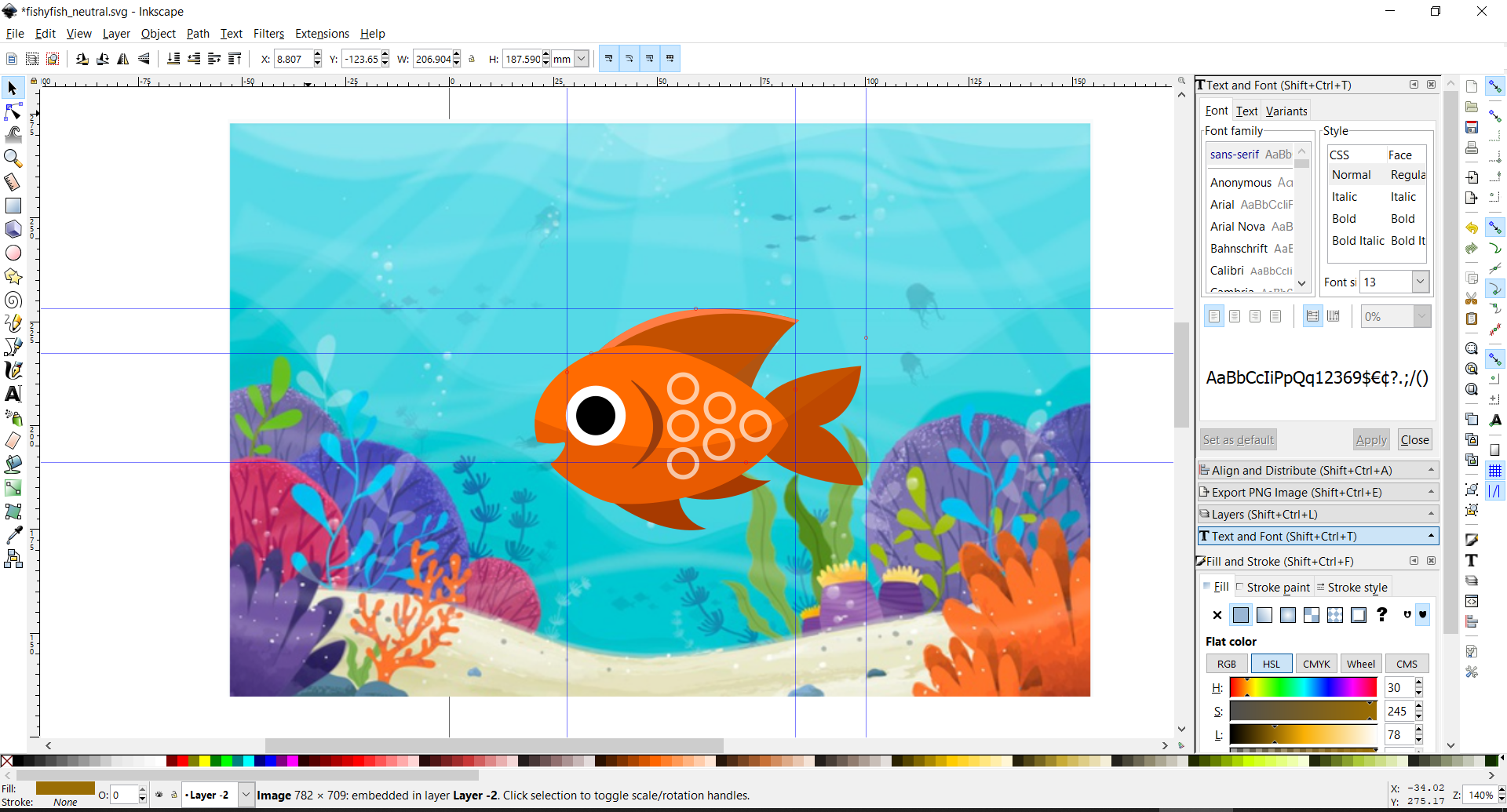
As we had already done the research into how to approach this issue, we decided to concatenate the two NLP models to see if we could create an “NLP Super Model” and, at the same time show a proof of concept for the concatenation of multimodal models. The approach and results can be read in section 5.2. “Concatenation of models”. The final model does not implement this concatenated model for reasons that we will explain in section 5.2.

*Addendum - Potential Regulatory Issues for Privacy and Transparency*

This addendum is technically not part of this Intelligent Interfaces project but was undertaken in the context of a course on an AI Ethics and Regulation course. It is attached to this document as already from the beginning we made comments that an application as our Goldfish of Solitude has serious privacy concerns. The addendum (that can be read in section 7 of this report) contains a legal analysis of emotion detection on Smart Home Assistants. Since our Goldfish can be classified as such an application, the legal analysis is entirely applicable.

# 3. Design of The Interface

One of the first steps was conceptualizing the interface. To do this, different options were explored. A Golden fish was chosen, inspired by the song Goldfish of Solitude. The golden fish would exist inside of a world (the aquarium), similarly to pet one could own. After there was a decision on the avatar, the fish design was started. It was established that the fish should have “sweet/cute” appearance so it would encourage interactions, and affection from the part of the individual, as a normal pet would. Visually, this “cuteness” of the fish was achieved through the use of soft shapes, bright colours and a cartoon-like appearance.



*Figure 1 – The Interface*

This cartoon aesthetic, using simple shapes, is not only more adequate for the elderly as it is easier for them to perceive but also fits well within the minimalistic UI design trends (such as: flat design, material design and simplistic geometry) that have been coming to trend over the past few years.

Material design was created by Google in 2014, with the aim of making the brand’s products visually coherent. This design style is defined by the following characteristics:

* Card-inspired interfaces, with multiple geometric spaces;
* Softened borders;
* Empty spaces;
* Two main fonts: Roboto and Noto;
* Simple and Quick animations;
* Responsive design;
* Shape, proximity and colors define the relationship between different objects;
* Visual hierarchy between objects;
* Use of Shadows.

Flat design was created in 2011 influenced by the Swiss Style of the 1920’s and it lives of the notion that less is more. It aims to fight the over-visual stimulation derived from the constant presence of technology in an individual’s life by simplifying interfaces. Flat design can be described by its following traits:

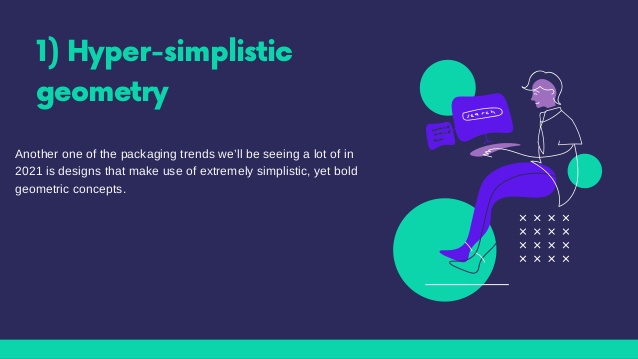
* Minimalist;
* Rejects gradients, shadows, heavy textures and glow;
* Vibrant, saturated Tones;
* Easy to read;
* High contrast between elements;
* Uses grids and geometric elements;
* Functionality-inclined;
* Bold and other easy-to-read fonts;
* Visually balanced.



*Figure 2 – Material Design and Flat Design*

Finally, there is Hyper-Simplistic Geometry. This is a design trend that gained traction in 2021. Hyper-Simplistic Geometry borrows elements from the previously presented design styles, incorporating the following aspects:

* Simple geometric shapes;
* Abstract elements;
* Vibrant tones;
* Motion-Inspired shapes;
* Overlap of visual elements
* Traced forms (as drawn shapes);
* Textures created by geometric forms;
* No shadows or glows
* Sharp angles.

****

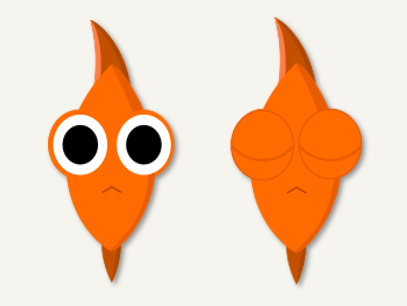
*Figure 3 – Hyper-Simplistic Geometry Design*

Regarding the color palette, we opted by choosing bright tones with high contrast. The first, in order to captivate the user and create an attractive interface. The second, to increase readability, an important factor to consider when designing for the elderly.

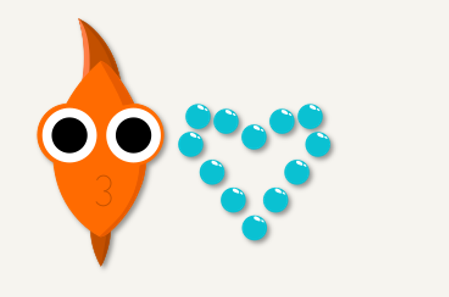


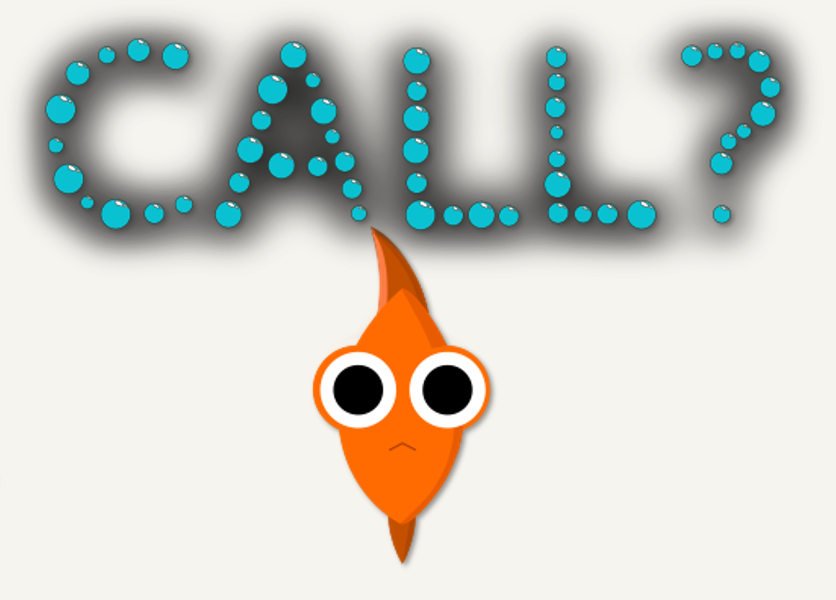
*Figure 3 – Color Palette*

The fish’s standard state consists of swimming around moving is tail in an animation consisting of five frames. Upon the detection of the emotion, it responds to the user visually. The standard emotion display of the fish is happiness. If the fish detects anger, it will open his eyes widely and blink twice.



*Figure 2 – Facing the User Figures 3 and 4 – Fear frames*

If the fish detects the user is sad, it will blow heart shaped bubbles, as in to comfort the user. If it detects fear, a set of bubbles will appear on top of the fish, together with the question “Call?”, if the answer is yes, the system will call the contact of someone (previously defined) close to the user, here we included a drop shadow, in order to increase readability on top of the blue background. Finally, if the fish detects joy it will keep it’s standard expression of happiness.

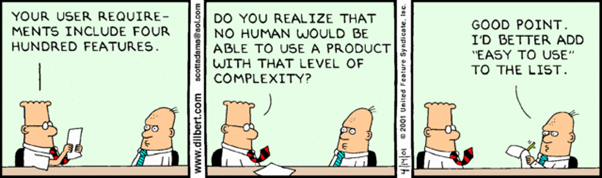
 *Figure 5 – Reaction to Sadness Detection Figures 6 – Reaction to Fear Detection*

Regarding the fishbowl where the fish lives this was decided after a poll in a public Facebook group with a membership of 28 thousand members in which two backgrounds were presented, the one on the screen and a simpler one that we considered that could be good in terms of readability. The colorful one won, with the argument of it being more visually/rich and potentially more engaging for elderly people.

Our interface was created with Inkscape (an open-source vectorization program, similar to illustrator) and Unity. Both the background image and the shape and color of the fish were chosen based on input from professionals working with the elderly, with expertise in the areas of Alzheimer and dementia. We have chosen for a colorful, stimulating background and the shape and color of the fish have been chosen for optimal likability (round shape, big eyes; mimicking baby mammal features) and color contrast with the background (maximizing visibility and resolution).

# Usability

In order to further understand the requirements of an intuitive interface, the work of Fogg, an important author in the fields of UI/UX, was explored in the development of this system. Making the interaction easy and intuitive is an imperative when designing an interactive interface. Even in situations where motivation is present, this alone is not enough to influence an individual’s behavior. The skill/capacity levels of an individual play a crucial role in the possibility of performing a certain behavior. This is particularly true for elderly people, who are often not as “fluent” in the digital world as the newer digital-native generations. According to Fogg (2009) “In order for behaviour to be occur, people must have some non-zero level of both motivation and ability”( pp.3). As such, increasing the user's ability/capacity of accomplishing the task, simplifying the situation, is an important factor to increase the probability of the behavior being performed.



*Figure X – Simplicity in Design*

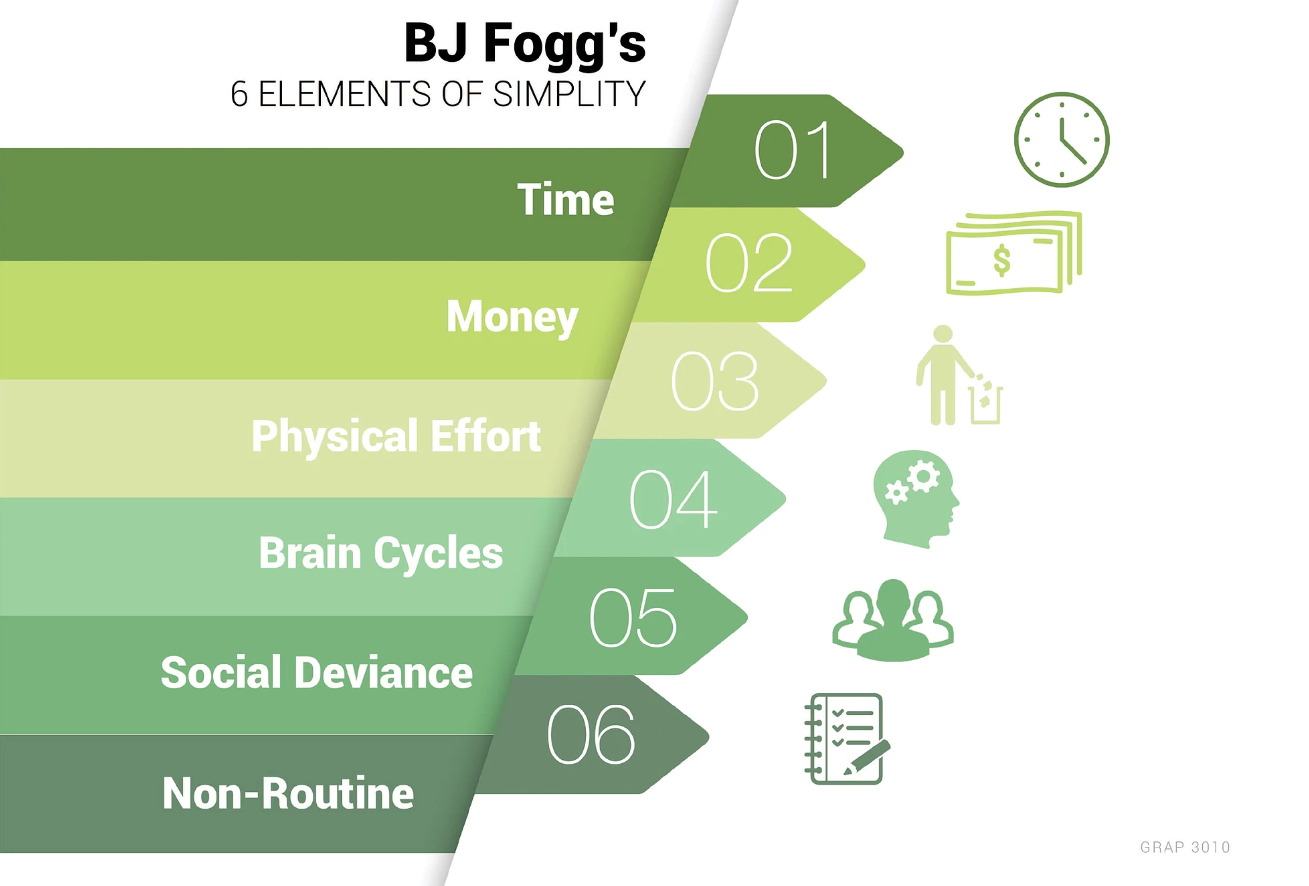
Fogg elaborates on the interaction between motivation and ability explaining that people with a low level of motivation can potentially carry out a certain behavior if it requires extremely low values ​​in terms of ability/capacity. The author illustrates this by giving the example of buying a car. Foog argues that, even though he does not need a new vehicle (having a low level of motivation), he would accept to buy one if it was sold to him for 1 dollar, as it is easy for him (having a high level of ability) to buy a car that is within this value (Fogg, 2009).

Fogg (2009) also presents a series of aspects that are important to consider for designing an interface that the user whishes to interact with, which he calls *Elements of simplicity*. These consist of six concepts influence an individual's ability to perform a certain task. Out of these, Brain Cycles was one of the concepts that was considered most relevant to this project. To further clarify, Brain Cycles is a factor that is associated with the mental demand that a certain task has. If this demand is high, the behavior will not be simple. Fogg explains that this happens because most of the time our mind is absorbed in other matters, and users often have little mental availability to get involved in a deep and complex way in the accomplishment of a task (Fogg, 2009). In elderly people, who can have health concerns such as dementia, this is an even more relevant matter. Considering this, we focused on decreasing the mental demand of the interaction with the fish avatar.

Another important *Element of Simplicity* explored for this interface was physical effort. The author states that elements that require it are not simple, illustrating with the example of someone who wants to walk from Las Vegas to Stanford. Such behavior would not be simple as it requires a lot in physical terms. When catching a train, this behavior would become simpler as it would not be as heavy in physical terms (Fogg, 2009). Once again, for elderly people, who at times might have less control over their movements and get tired more easily, this factor is even more relevant. Here, systems that allow for voice-interaction, such as the created fish, are preferable in the sense that they require less effort than other systems.

Finally, there is the *Element of Simplicity* Non-Routine, also relevant to the project. This element stems from the fact that people find it easier to implement behaviors consistently when they are part of their routine. In other words, behaviors that diverge from routine may be seen as more complex (Fogg, 2009). In this case, by allowing the user to talk to the fish as they would to a pet, or even a plant, there is less deviation from the behavior that the average person could carry out. This way, the system keeps the individuals “company” while monitoring their well-being in a fairly natural and intuitive manner.

Regarding other interactive elements, computer systems are generally poorly adapted to the user, attempting to stimulate him in a frustrating way. This is characterized by interfaces being tedious and inconvenient for the user, which distances him from the possibility of achieving the desired behavior. “Computer systems often do a frustrating job of triggering behavior. Spam, pop-ups ads, and other annoying artifacts are actually triggers” (Fogg, 2009, pp.3). In this project this is achieved by the intuitive interface, easy to interact with through voice, avoiding icons, pop-ups, further navigation items and unnecessary calls-to action that could confuse the user. All these aspects also contribute to a higher level of ability from the user’s part, making it more likely that they choose to interact with the system, and integrate it into their lives.



*Figure x – Fogg’s 6 elements of simplicity*

# 5. Emotion detection

With our project we would like to be able to report on a variety of emotional states. While researching our project we quickly discovered that there are not that many existing projects on emotion detection as we might have hoped for. Most projects deal with sentiment analysis only. Sentiment analysis discerns only three “sentiments”: positive, negative, and neutral. Polarity analysis has even fewer dimensions, as it only deals with positive or negative polarity. Unfortunately, annotated data sets reflect this bias. There are many data sets annotated for sentiment analysis or polarity, but not very many data sets annotated for multiple emotions.

Emotion detection is based on Emotion Models that define how emotions are represented[[1]](#footnote-1). For emotion detection usually discrete and dimensional models are used. Choosing to represent emotions discretely in stead of on a continuum has unavoidable consequences for the correctness of the analysis. We choose to lump together various emotional states.

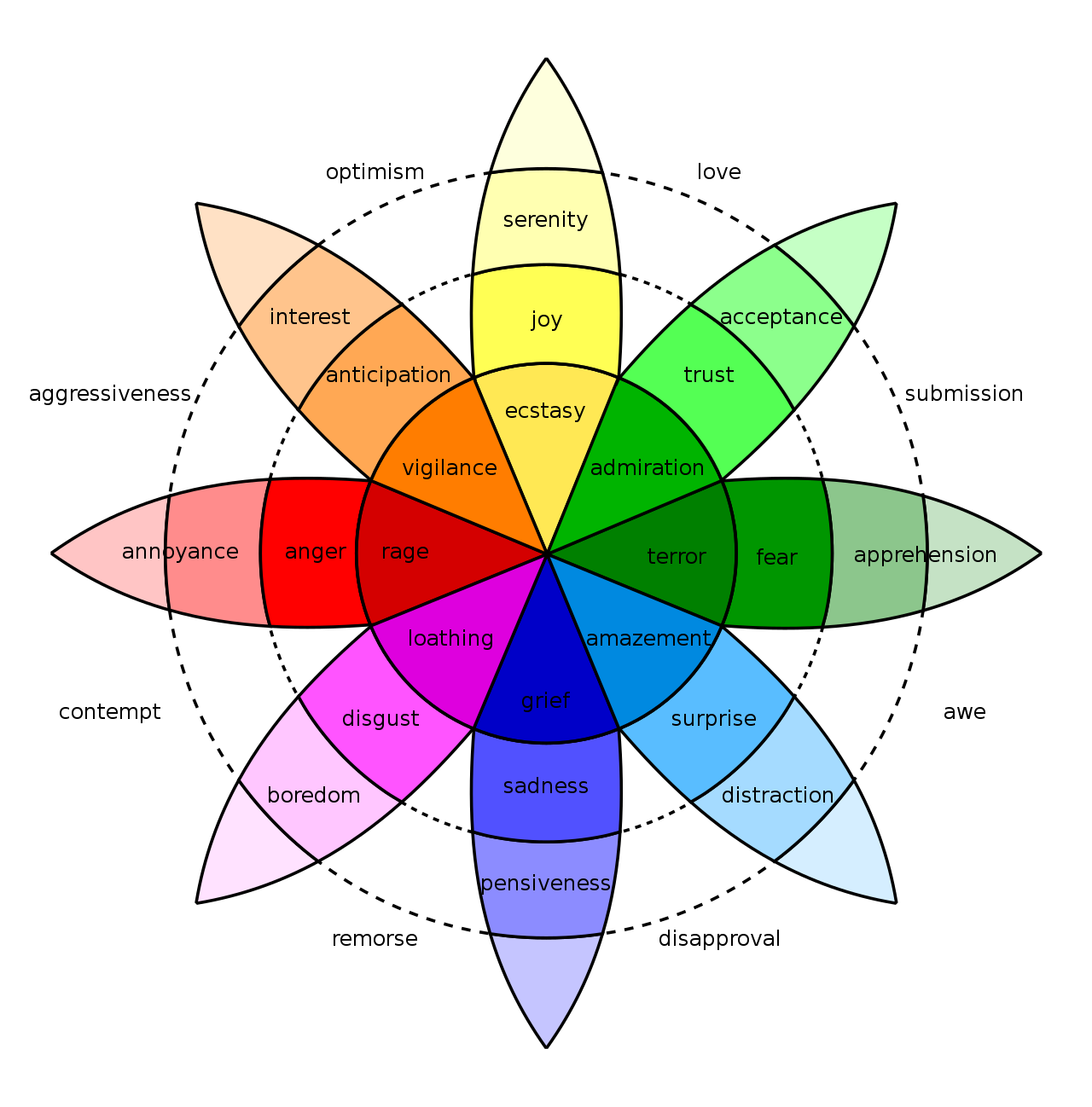
* 1. Universal -human- expressions (Ekman)

With respect to the discrete emotion models the model of US psychologist Paul Ekman is often used. In the 1960s and 1970s Ekman suggested that emotional expressions are universal. His model distinguishes emotions into six basic categories or Fundamental Emotions: happiness, sadness, anger, disgust, surprise, and fear (contempt was added at a later stage)[[2]](#footnote-2). Ekman derives these fundamental emotions from facial expressions. Interestingly, Ekman offers training tools for humans to learn how “read the hidden emotions of those around you”. Indicating that although he postulates that the expressions of human emotions may be universal, humans may not be all that good in interpreting these expressions of emotions. A sentiment that is echoed by Aleix Martines, a researcher at Ohio State University (Columbus), who researches this topic.

Ekman’s work is being challenged by psychologists and cognitive scientists, who believe that facial expressions vary widely between contexts and cultures. The journal Psychological Science in the Public Interest asked a panel to review the literature, which ended up being at least a thousand papers. They concluded that there is little or no evidence that people can reliably infer someone else’s emotions from their facial features. Humans need more features to derive emotional states than just facial features alone. Body movement, knowledge of someone’s personality, tone of voice, changes in skin tone, even context are important features [[3]](#footnote-3).

* 1. Wheel of Emotions (Plutchick)

Plutchick’s[[4]](#footnote-4) Wheel of Emotions[[5]](#footnote-5) (illustrated in Figure 3), is also an often-cited theoretical framework for discrete emotion classification.



*Figure 3 - Plutchick’s Wheel of Emotions*

The wheel is divided into (8) primary emotions, whereby each primary emotion has a polar opposite:

**Primary**: The eight sectors are designed to indicate that there are eight primary emotions: anger, anticipation, joy, trust, fear, surprise, sadness, and disgust.

**Opposites**: Each primary emotion has a polar opposite. These are based on the physiological reaction each emotion creates in animals.

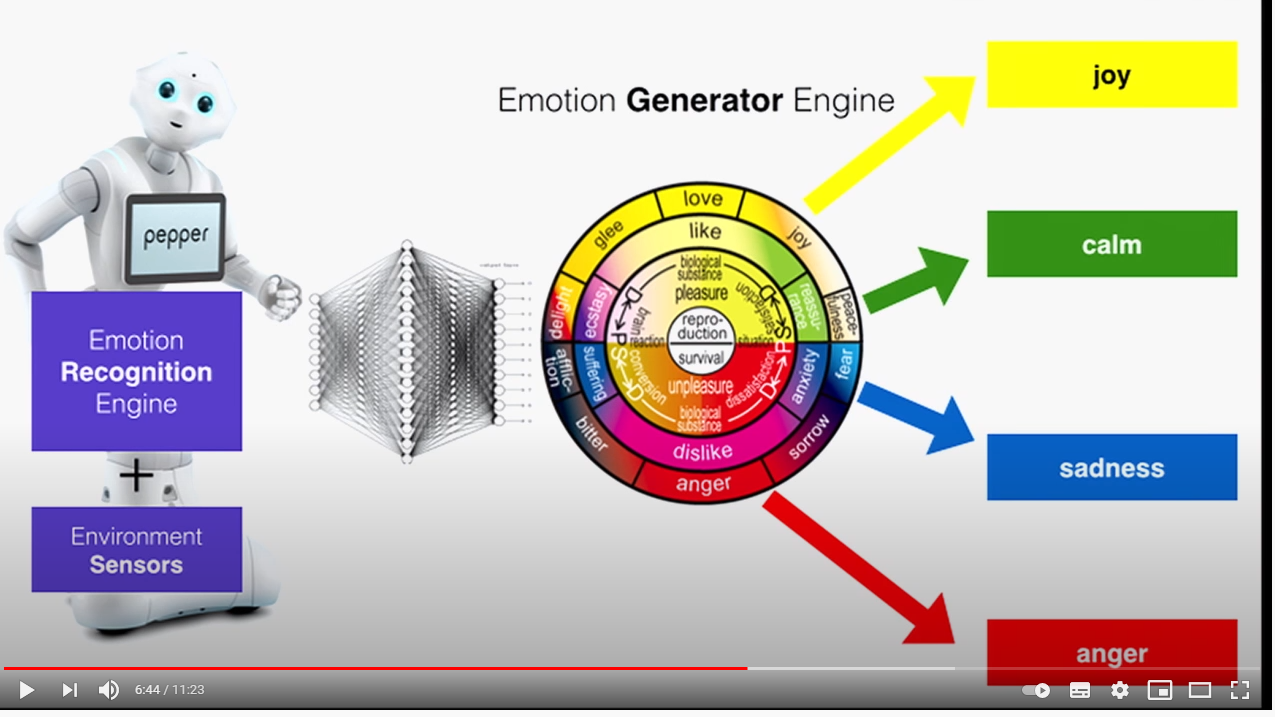
* Joy is the opposite of sadness. Physiology: Connect vs withdraw
* Fear is the opposite of anger. Physiology: Get small and hide vs get big and loud
* Anticipation is the opposite of surprise. Physiology: Examine closely vs jump back
* Disgust is the opposite of trust. Physiology: Reject vs embrace

The Wheel of Emotions has more granularity and the emotions listed are more diverse and have modalities than Ekman’s model, but as this emotion wheel was based on animal behaviors. The emotionality features are no longer “read” from specific facial features, but from signals from the entire body.

* 1. More granularity and multimodality

Orthony, Clore, and Collins (OCC) discretised 22 emotions, 16 additional emotions to Ekman’s Basic Emotions, such as relief, envy, reproach, self‐reproach, appreciation, shame, pity, disappointment, admiration, hope, fears‐confirmed, grief, gratification, gloating, like, and dislike. The OCC model distinguishes emotions involving a focus on events, on actions and on objects. Emotions involving outcomes of events are distinguished by whether they are one’s own (e.g., sad) or another’s outcomes (e.g., pity). They also differentiate prospective outcomes (e.g., fear) from known outcomes (e.g., grief). Not all emotions are about events. Some emotions are about the agency of actions and are about whether actions are praiseworthy (e.g., pride) or blameworthy (e.g., shame), etc.[[6]](#footnote-6) In other words: Emotions do not exist without context and when humans interpret emotions, they will assess much more than just facial features.

A similar, yet different emotional complexity is shown by the model presented by Mitsuyoshi which was used for SoftBank’s Pepper robot’s Emotion Recognition Engine.



*Figure 4 - Pepper’s Emotion Generator Engine[[7]](#footnote-7)*

Unfortunately, the model is in Japanese, so it is difficult for us to draw serious conclusions, but we can see the multi-dimensionality of what is finally distilled into emotions (see Figure x above).



*Figure 5 - Mapping of emotions from a recording of a motorcycle ride (Mitsuyoshi)*

Mitsuyoshi uses a multimodal mapping of emotions that is not accessible to us for our project, as he uses functional Magnetic Resonance Imaging to map emotions during conversation. This is a modus which is not even accessible to humans.



*Figure 6 - Mapping of emotions from brain activity (Mitsuyoshi)*

Apparently developing the Pepper emotion engine was difficult, because Pepper sold to businesses in 2015 with the emotion engine turned off[[8]](#footnote-8).

* 1. Conclusions

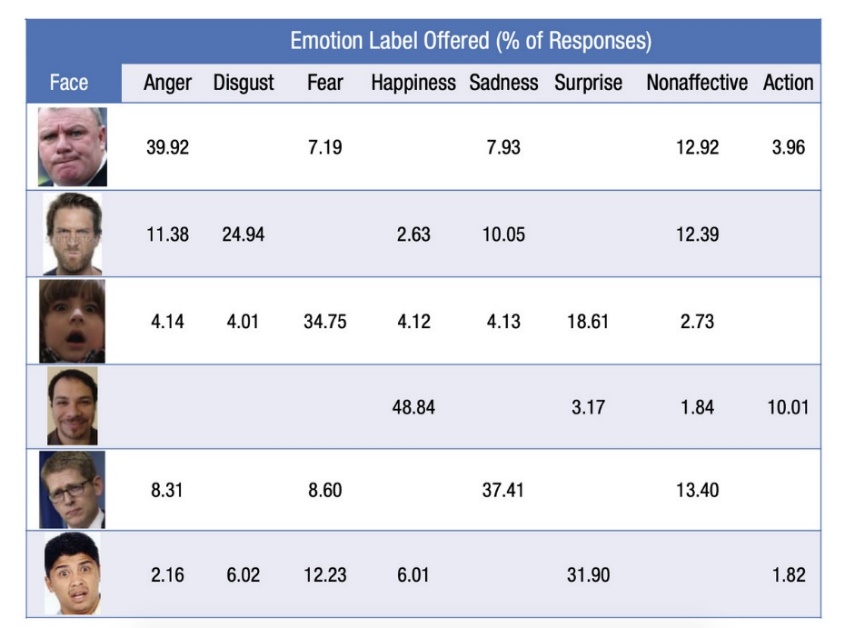
In view of the above it seems that any selection of emotion features will be a flawed one whereby the more complex models are well beyond the scope of this project. This project takes as the emotional target features Eckman’s six basic emotions: happiness, sadness, anger, disgust, surprise, and fear. With a caveat for “disgust,” as some of our data sets do not cater for this emotion and the confusion with other emotions is also great.

Nevertheless, it has to be kept in mind from the theoretical background that:

1. The more discrete and the less granular the model, the less accurate it is in reading true human emotion.
2. Eckman’s model is derived from facial features alone, there is no multimodal input.
3. Humans are not very good at concluding emotions from facial features.

In our model choice, we have been partially led by the availability of annotated data sets. The most fully annotated data sets we have found all select six primary emotions, usually consisting of: happiness, sadness, anger, disgust, surprise, and fear, which are the Basic Emotions as described by Eckman.

It is clear that the way in which the feature set has been established may also have an impact on how data sets are labeled. When human annotators are asked to label data and they are offered a limited set of emotion labels, there will be reductive bias in the labeling.



*Figure 7 - When people are asked to label emotions on faces and aren’t given a set of choices, their answers vary considerably, as this chart shows. Image: Barrett et al.*[[9]](#footnote-9)

Any Machine Learning process applied to these labeled data sets will only be as good as the labeling allows. It is therefore quite important to make a thorough assessment of the data sets used: how have they been created, how have they been labeled, and what is the potential impact on the outcome of the project.

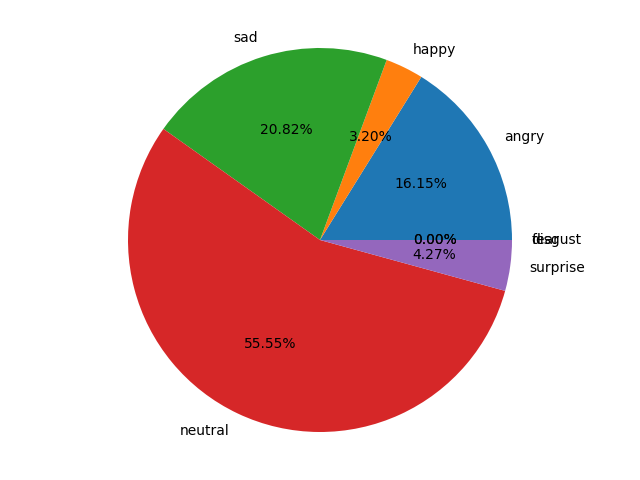
# Multimodality

In the section above it was clarified that humans are not very good at concluding emotions from facial features alone. We have at least facial, auditory, and textual (semantical, syntactical) inputs when we evaluate the emotional status of another person. Usually, we also have contextual clues and if we know the person, we also have information about their character and on how they usually behave. Humans process information multimodally.

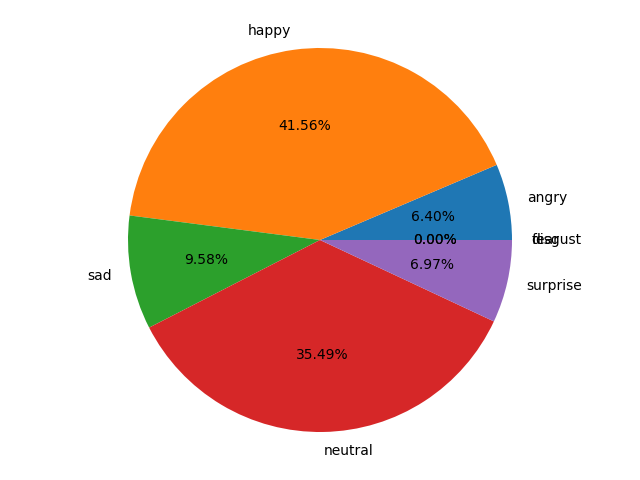
Whereas from the literature, it is clear that facial expression is insufficient to conclude which emotion applies to a particular expression, it should be clear that the same is valid for semantical and syntactical expression. To demonstrate this fact, we have conducted an experiment.

We recorded videos of a person saying the sentence “Yes, I am very happy for you”, while expressing different emotions. We have then put these videos through our preliminary vision system to demonstrate that the emotion that is detected by the visual system is different, depending on how the sentence is uttered:

*Angry*

*Happy*

When a person says “Yes, I am very happy for you” with an open, smiling face, we can assume that they are truly happy for us. When the emotion expressed is anger, we will not assume that they are happy for us. The meaning of the words will become secondary, and the emotion as expressed though prosody and (here) visual clues will become the primary clues that will put a spin on the meaning of the words. From the above, it can be concluded that text may not be the best way to convey or analyze emotion as text may not convey sufficient emotional cues. This is why it is potentially better to work multimodally for our project.

Multimodality however brings its own challenges, especially with respect to the data sets. Emotionally annotated multimodal data sets are usually annotated based on the full set of multimodal inputs. It is impossible to know which input modality provided the clue for the mark-up of the emotion. Was it the facial features, the tone of voice or was it what was said that was considered the reason for annotating a certain emotion? Potentially there were even factors in play such as the situation or the context, none of which would be captured by our system.

For NLP the problematic situation with the availability of data sets may even be exacerbated by the fact that there is a big difference between written language and spoken language. Our project involves spoken language that is converted to text before it is analyzed for emotionality. Most research in the field of textual emotion detection seems to have been done on written text. The latter matter is an issue when looking for data sets for the NLP part of the project. Most data sets that are annotated for emotionality concern written text, not spoken language. It is difficult to find emotionally annotated sets for spoken language.

With respect to the model, there is the issue of how to bring together the three modalities. Do we allocate weights to each of the models before we concatenate them? If yes, how do we assign the weights? If we do not add weights, but otherwise bring the models together, how do we do this? By using an ensemble method with voting procedure. By vectorizing the outputs of the respective models and concatenating the vectors “as is” and putting them through another model? Does the order in which we put the individual models together matter? Is it perhaps better to first compare the models two by two to see which binary combination gives the best result? These are all issues that a multimodal model generates. Multimodality could potentially improve the performance of our emotion detection, but the issues thrown up by the added complexity are numerous.

**Additional documentation on multimodality:**

French multimodal project (with unfinished paper)

<https://github.com/maelfabien/Multimodal-Emotion-Recognition/blob/master/README.md#viii-deployment>

**Multimodal data sets:**

* Overview pages
  + <https://github.com/declare-lab/awesome-emotion-recognition-in-conversations>
  + <https://github.com/A2Zadeh/CMU-MultimodalSDK?fbclid=IwAR2j0eOaVR-kMlOATWtUmBTdpgjrfOSziwF7GbeXLYPIhNJmOkOGAYe6TSM>
  + <https://huggingface.co/datasets/daily_dialog>

**Actual data set locations:**

<https://github.com/SenticNet/MELD> (MELD) => **implemented**

Obtained from dialogues and utterances in a Television Show called Friends.

**Way to create own data set**

Data sets: multimodal vs non-multimodal, own data set YouTube videos with transcripts (not annotated)

<https://pypi.org/project/youtube-transcript-api/>

**Multimodal concatenation:**

<https://paperswithcode.com/sota/multimodal-sentiment-analysis-on-cmu-mosei-1>

Ensemble Methods:

<https://www.youtube.com/watch?v=MHqvWPw3BVA>

# System 1

## 7.1 The individual models

This section describes System 1 mostly from the point of view of the NLP modality. Comments in the sections on video and audio input are purely from shared experiences that touch upon the NLP modality in some way.

## Input from camera (video)

From our brief multimodiality experiment with video and sentences that are exactly the same syntactically, but that can be expressed with a different emotion to gain an entirely different meaning semantically, we have concluded that to capture emotion the sampling rate of the video stills should be sufficiently high. Faces are not expressing emotion rigidly. They change fast, even when pronouncing only a brief sentence. Also, averaging out emotion scores over the period of an entire utterance will lead to scores averaging out to “neutral, as the face seems to be in its neutral state a lot of the time, even as it transitions through various emotionally expressive positions. This might be an issue technically. The human visual system can process 10 to 12 images per second and perceive them individually, higher rates are perceived as motion. This means that human expressions and the interpretation thereof by humans is based on this physical reality. Do humans interpret facial expressions as “stills” or do they interpret minute motions (from one position to another at a very rapid rate of approximately 15 samples per second) of the facial muscles. And if the latter, how slowly do they need to change in order for a human to detect the change and the accompanying emotion, and more importantly is our camera able to sample at this rate, and if it is able to sample at this rate, is our hardware sufficient to process all the images sufficiently rapidly.

From stills it has been researched that humans can recognize happiness in 23 to 28 ms, while neutral, disgust and surprise take 3 to 4 times longer as does fear. Sadness and anger are recognized 10 times slower than happiness[[10]](#footnote-10).

## Input from microphone (audio)

Google Speech to Text Cloud module transposes Speech into text for the two NLP models.

## Input from language (NLP based on STT from transformed audio input)

1. **Emotion detection from text:**

Whereas for detecting emotions from voice/speech, images, and other multimodal methods there exists an exhaustive knowledge base, there exists great paucity in research for texts. This is because unlike multimodal methods, texts may not portray peculiar cues to emotions.[[11]](#footnote-11) This was more or less confirmed by our little experiment.

1. **Data sets for NLP Emotion detection**

<https://www.kaggle.com/shrivastava/isears‐dataset> (ISEAR) => **implemented**

(Discrete) Obtained from cross‐cultural studies in 37 countries and contains 7665 sentences annotated for joy, sadness, fear, anger, guilt, disgust and shame.

<https://www.aclweb.org/anthology/I17‐1099/> (DailyDialog)

Contains 13118 Dialogues extracted from conversations and annotated for happiness, sadness, anger, disgust, fear, surprise, and others

<https://www.crowdflower.com/wp‐content/uploads/2016/07/text_emotion.csv> (CrowdFlower)

Constructed from 39,740 tweets and annotated for thirteen(13) emotions

<http://alt.qcri.org/semeval2017/task4/index.php?id=download‐the‐full‐training‐data‐for‐semeval‐2017‐task‐4> (SemEval-2028 Task 4) => does not seem appropriate for our goal

(Discrete) Data contains 1250 texts obtained from Tweets, News headlines, Google News and other major newspapers. Annotated for Ekman's 6 basic emotions.

[http://people.rc.rit.edu/∼coagla/affectdata/index.html](http://people.rc.rit.edu/~coagla/affectdata/index.html) (Cecilia Ovesdotter Alm's Affect data) => does not seem appropriate for our purposes

Constructed from Tales and classified into angry, fearful, happy, sad, disgusted and surprised emotions.

1. **Models**

From the three approaches below the Machine Learning approach was chosen for Model 1, as this is the model most familiar from what was covered in the postgraduate so far.

* Rule construction approach
* Machine Learning approach
* Hybrid approach

Google BERT is touted as a state-of the art pre-trained basis for NLP oriented work. It is also fully implemented on the Google Colab environment and therefore Google BERT was first choice as a basis for our Model 1.

The NLP2 notebook used for this project consisted of an LSTM + CNN algorithm. This model comprises a Bidirectional LSTM (a type of Recurrent Neural Network) with a CNN (Convolutional Neural Network) layer. While the first one (the LSTM Layer) grabs the information about the context of sentences, the second one (the CNN) extracts local features.

1. **NLP Model 1: BERT + Hugging Face Transformer Library**

The first NLP model is a model based on BERT[[12]](#footnote-12)[[13]](#footnote-13) (pre-trained language embedding) in combination with the Hugging Face library, which appears to be the most widely accepted Pytorch interface for working with BERT.

Tokenization is performed with a BERT-specific tokenizer (from transformers import BertTokenizer). The BertTokenizer creates a token for every item in an utterance. This means that there are also tokens for punctuation marks (see example below). The tokenizer also adds a token at the beginning of every utterance ([CLR]) and a separator token at the end of every utterance ([SEP]). The Token ID’s are numbers that refer to the line item number in the BERT Lexicon for the word, partial word, letter, or special token. The word “our” will always be tagged with Token ID 2256, for instance.

Original: Our friends won't buy this analysis, let alone the next one we propose.

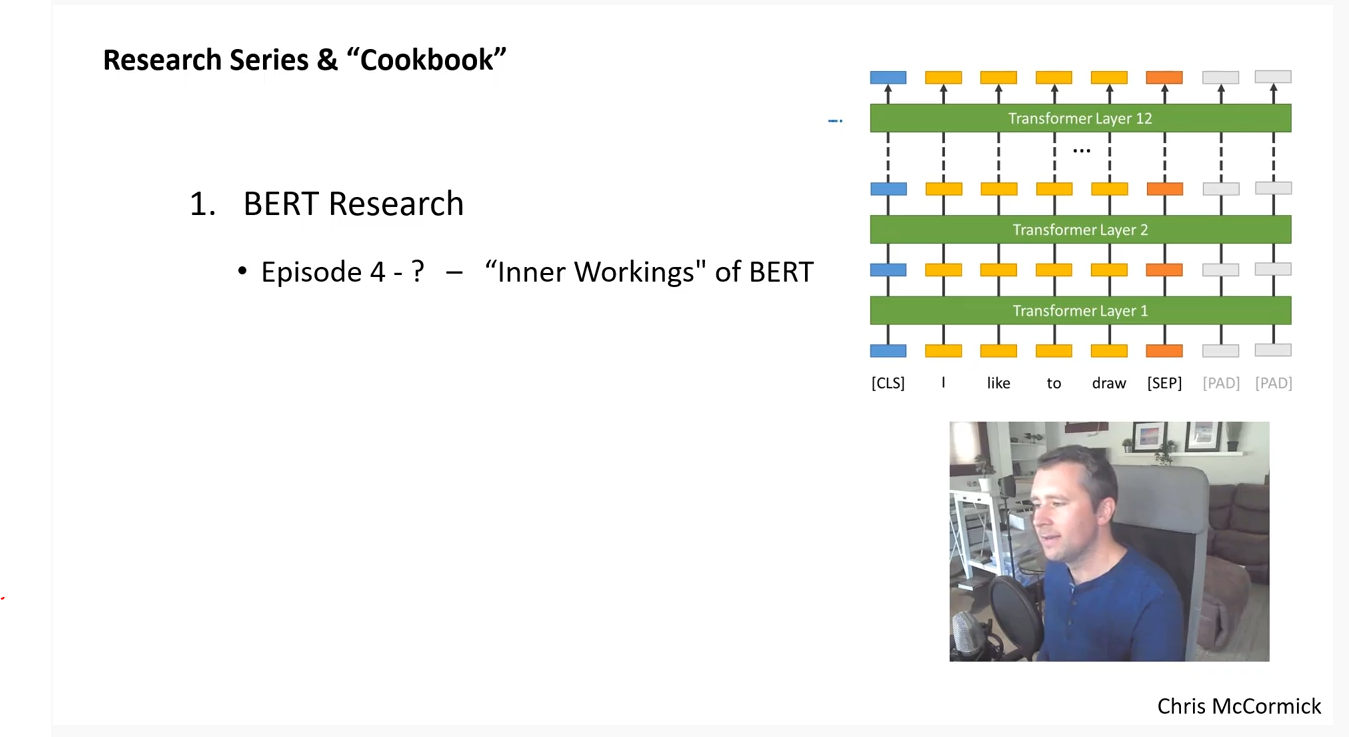
Tokenized: ['our', 'friends', 'won', "'", 't', 'buy', 'this', 'analysis', ',', 'let', 'alone', 'the', 'next', 'one', 'we', 'propose', '.']

Token IDs: [2256, 2814, 2180, 1005, 1056, 4965, 2023, 4106, 1010, 2292, 2894, 1996, 2279, 2028, 2057, 16599, 1012]

The {SEP] token is a remnant of one of the pre-training tasks that BERT was trained on to learn how to embed language, namely a comparison task between two sentences (does sentence x precede sentence y). Even if we only feed BERT a single utterance, the [SEP] token is still required.

The [CLS] token must be used at the beginning of an utterance for all classification tasks. “CLS” is short for “classification” or “classifier”. This token is special.

BERT base (the BERT used in Model 1) has 12 Transformer (encoder) layers (BERT large has 24)[[14]](#footnote-14). Each Transformer takes in a list of token embeddings and produces the same number of embeddings as output (with the features changed by the Transformers). At the output of the 12th transformer only the first embedding (designated by the [CLS] token) is used by the classifier. This token is used as the aggregate sequence representation for classifier tasks[[15]](#footnote-15).

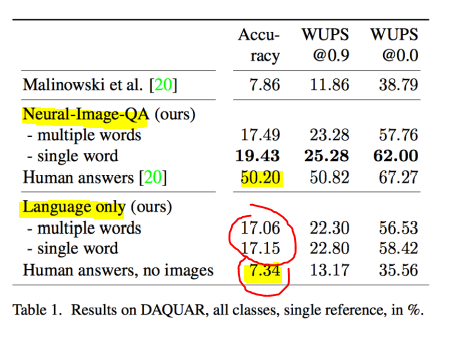


*Figure 8 – Illustration of the BERT tokenization and Transformer layers[[16]](#footnote-16)*

The BERT model is then fine tuned to the particular task at hand. In our case this is a multi-class classifier that classifies text towards multiple emotions.

An interesting, yet disturbing fact we have learned about BERT and other huge pre-trained models is that it is unclear how much the pre-trained models have already learned about the world (and language) before we fine tune the model with the actual task we would like it to perform. An interesting example is given here by the results of a Question Answering model by Malinowski[[17]](#footnote-17) et al. Whereby questions were asked about images (e.g. “how many drawers are there in the chest?”). The language portion of the model was based on a pre-trained BERT model. When the question was asked in the presence of the image in question, humans scored an accuracy of approximately 50%, whereas the highest score (single word) of the machine model was only around 19%. Oddly enough, when the image was not shown, and humans had to revert to guessing what was in the picture their score dropped to around 7% whereas the machine model’s score remained virtually the same (it only dropped by 2 percentage points to 17%).

It has to be concluded that the pre-training somehow conveys information to the system about the world, a kind of pre-conceived notion. It is therefore difficult to ascertain in this case how much information the model that was fine-tuned towards computer vision actually contributes to the final answer that is provided by the fine-tuned system. Is the complete model doing much computer vision when it answers questions about images, or is it mainly reverting to what the BERT model has learned (in this case how many drawers does a chest have, on average, in texts describing chests and drawers).



*Figure 9 - ?*

It is distinctly possible that by using the pre-trained BERT model, we are introducing a similar pre-learning bias and the system will learn less from our fine-tuning (especially if the data set is small) than we might think. The consequence hereof is that the model may not actually evaluate the emotions in the text that is used to fine-tune the model, but that it is somehow giving what it knows to be “average” emotions. This could be investigated by applying attention models, that look at where the attention is going, but this is well beyond the scope of this project and therefore this fact is merely noted but not acted upon in the course of this project.

**BERT background information**

Explanation of BERT (McCormick)

<https://www.youtube.com/channel/UCoRX98PLOsaN8PtekB9kWrw/videos>

BERT Fine-Tuning Tutorial with PyTorch (McCormick)

<https://colab.research.google.com/drive/1Y4o3jh3ZH70tl6mCd76vz_IxX23biCPP#scrollTo=86C9objaKu8f>

Smart batching tutorial with BERT (McCormick)

<https://colab.research.google.com/drive/1Er23iD96x_SzmRG8md1kVggbmz0su_Q5>

**BertForSequenceClassification**

<https://huggingface.co/transformers/v2.2.0/model_doc/bert.html#bertforsequenceclassification>

Classify Emotions in text with BERT

<https://www.kaggle.com/praveengovi/classify-emotions-in-text-with-bert>

1. **BERT model finetuned with a classifier on the MELD (multimodal) data set.**

The MELD data set is a multimodal, annotated data set obtained from dialogues and utterances from the television show “Friends”.

**Number of sentences: 1,109**

This is a very small set to use for fine tuning and may therefore not be appropriate, as BERT requires around 10 000 exemplars for fine-tuning. It was however the only multimodal data set for which we had access to emotion annotated sentences and accessible as a csv file.

**The unique emotions, as annotated: 'sadness', 'surprise', 'neutral', 'joy', 'anger', 'disgust', 'fear'**

These emotion categories are sufficiently varied.

**The data set is not balanced at all.**

Emotion emotion\_num

Sr No.

1 sadness 0

2 surprise 1

3 neutral 2

4 joy 3

14 anger 4

74 disgust 5

103 fear 6



The fact that the data set is so unbalanced is a second reason why it may not be ideal for purpose.

**For tokenization BertTokenizer was used from bert-base-uncased**

Uncased was chosen because the input will be from STT and there will be no capitalization of relevance to inform the training.

**After tokenization the train/test split was made with 80% training, 20% testing, random state 500.**

**Batch size was set to 16 (BERT creators recommend a batch size of 16 or 32 for finetuning BERT on a specific task)**

**The optimizer that was chosen was AdamW with learning rate 2e-5 and epsilon value 1e-8**

An opmizer optimizes the loss function.

We have chosen AdamW in stead of simple Stochastic Gradient Descent because the training is much faster with AdamW and AdamW has been significantly upgraded from its predecessor Adam. With Adam, the weight decay was implicitly bound to the learning rate. This means that when optimizing the learning rate a new optimal weight decay for each learning rate should be found. The AdamW optimizer decouples the weight decay from the optimization step. This means that the weight decay and learning rate can be optimized separately, i.e. changing the learning rate does not change the optimal weight decay. This allows for improved generalization performance.

We have set the learning rate very small in view of the small size of the data set and we have set the epsilon value to the value recommended in the literature.

**Number of training epochs was set to 5**

The recommended number of epochs is between 3 and 5.

**Below are the figures for the five epochs (accuracy, precision, recall and F1-score)**

======== Epoch 1 / 5 ========

Training...

Batch 40 of 56. Elapsed: 0:00:05.

Average training loss: 1.66

Training epoch took: 0:00:07

Running Validation...

Accuracy: 0.46

Validation took: 0:00:00

precision recall f1-score support

0 0.0000000 0.0000000 0.0000000 20

1 0.6666667 0.0769231 0.1379310 26

2 0.4675926 1.0000000 0.6372240 101

3 0.0000000 0.0000000 0.0000000 28

4 0.0000000 0.0000000 0.0000000 32

5 0.0000000 0.0000000 0.0000000 6

6 0.0000000 0.0000000 0.0000000 9

accuracy 0.4639640 222

macro avg 0.1620370 0.1538462 0.1107364 222

weighted avg 0.2908116 0.4639640 0.3060623 222

======== Epoch 2 / 5 ========

Training...

/usr/local/lib/python3.7/dist-packages/sklearn/metrics/\_classification.py:1272: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero\_division` parameter to control this behavior.

\_warn\_prf(average, modifier, msg\_start, len(result))

Batch 40 of 56. Elapsed: 0:00:05.

Average training loss: 1.40

Training epoch took: 0:00:07

Running Validation...

Accuracy: 0.57

Validation took: 0:00:00

precision recall f1-score support

0 0.0000000 0.0000000 0.0000000 20

1 0.5714286 0.4615385 0.5106383 26

2 0.6462585 0.9405941 0.7661290 101

3 0.5625000 0.3214286 0.4090909 28

4 0.2702703 0.3125000 0.2898551 32

5 0.0000000 0.0000000 0.0000000 6

6 0.0000000 0.0000000 0.0000000 9

accuracy 0.5675676 222

macro avg 0.2929225 0.2908659 0.2822448 222

weighted avg 0.4708464 0.5675676 0.5017366 222

======== Epoch 3 / 5 ========

Training...

Batch 40 of 56. Elapsed: 0:00:05.

Average training loss: 1.21

Training epoch took: 0:00:07

Running Validation...

Accuracy: 0.58

Validation took: 0:00:00

precision recall f1-score support

0 0.2000000 0.0500000 0.0800000 20

1 0.5200000 0.5000000 0.5098039 26

2 0.6643836 0.9603960 0.7854251 101

3 0.5294118 0.3214286 0.4000000 28

4 0.3103448 0.2812500 0.2950820 32

5 0.0000000 0.0000000 0.0000000 6

6 0.0000000 0.0000000 0.0000000 9

accuracy 0.5810811 222

macro avg 0.3177343 0.3018678 0.2957587 222

weighted avg 0.4926906 0.5810811 0.5172318 222

======== Epoch 4 / 5 ========

Training...

Batch 40 of 56. Elapsed: 0:00:05.

Average training loss: 1.04

Training epoch took: 0:00:07

Running Validation...

Accuracy: 0.60

Validation took: 0:00:00

precision recall f1-score support

0 0.5714286 0.2000000 0.2962963 20

1 0.6086957 0.5384615 0.5714286 26

2 0.6906475 0.9504950 0.8000000 101

3 0.5555556 0.3571429 0.4347826 28

4 0.2857143 0.3125000 0.2985075 32

5 0.0000000 0.0000000 0.0000000 6

6 0.0000000 0.0000000 0.0000000 9

accuracy 0.6036036 222

macro avg 0.3874345 0.3369428 0.3430021 222

weighted avg 0.5482363 0.6036036 0.5554469 222

======== Epoch 5 / 5 ========

Training...

Batch 40 of 56. Elapsed: 0:00:05.

Average training loss: 0.93

Training epoch took: 0:00:07

Running Validation...

Accuracy: 0.60

Validation took: 0:00:00

precision recall f1-score support

0 0.4444444 0.2000000 0.2758621 20

1 0.5833333 0.5384615 0.5600000 26

2 0.7045455 0.9207921 0.7982833 101

3 0.5714286 0.4285714 0.4897959 28

4 0.3055556 0.3437500 0.3235294 32

5 0.0000000 0.0000000 0.0000000 6

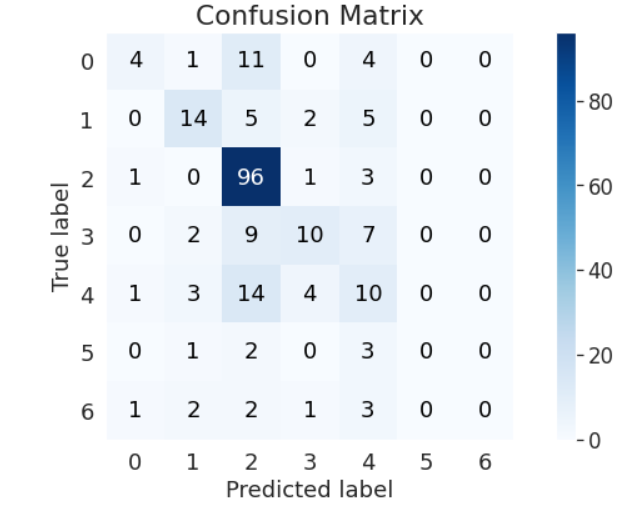
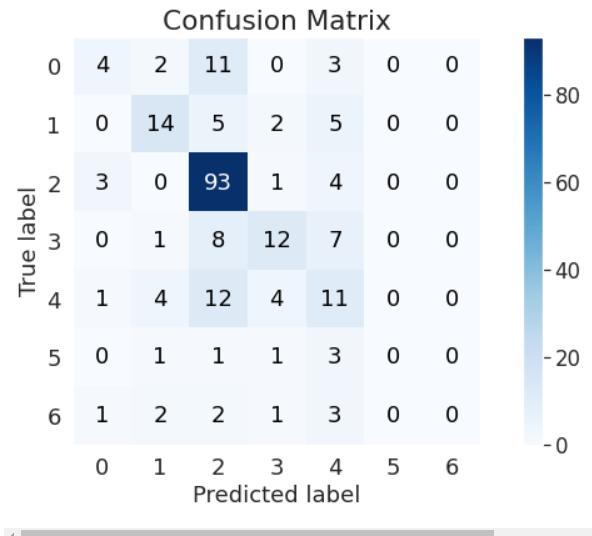
6 0.0000000 0.0000000 0.0000000 9

accuracy 0.6036036 222

macro avg 0.3727582 0.3473679 0.3496387 222

weighted avg 0.5450109 0.6036036 0.5620319 222

The performance of the model is fairly low with just a 60% accuracy.

The confusion matrices of Epochs 4 and 5 respectively (the best ones with 60% accuracy score) show that disgust and fear are never classified correctly, and anger is often misclassified as surprise.

Emotion emotion\_num

Sr No.

1 sadness 0

2 surprise 1

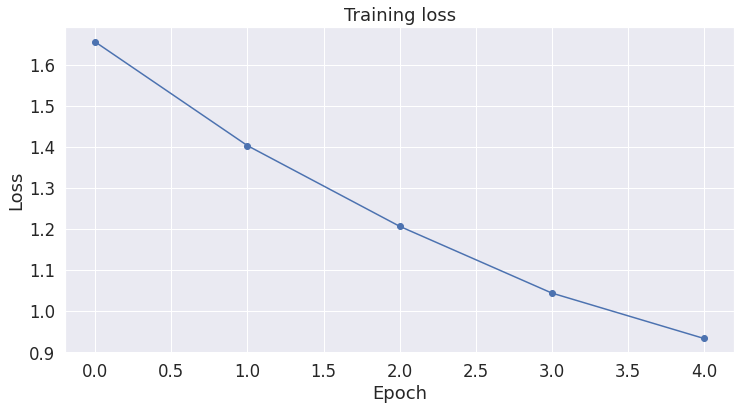
3 neutral 2

4 joy 3

14 anger 4

74 disgust 5

103 fear 6



After having changed some of the parameters such as 90% training and 10% test split, optimizer set to, learning rate (even smaller), epochs (up to 20) and not gaining any significant accuracy.

A renewed look at the data set provided an answer: the multimodal data set is actually annotated on the entire video input (sound, vision and text), and it is not at all the case that the annotation is made based on the textual content. This means that the annotations here are not “logical” and that they are often not applicable to the text at all (see our experiment in section 4 “Multimodality”).

The conclusion is that the MELD data set should not be used to train the NLP part of our project and a better data set should be found.

1. **BERT model finetuned with a classifier on the ISEAR data set**

Student respondents, were asked to report situations in which they had experienced all of 7 major emotions (joy, fear, anger, sadness, disgust, shame, and guilt). These data were collected into the ISEAR dataset.

**Number of sentences: 7666 but after cleaning badly coded rows, only 7503 remained**

This data set is quite a bit bigger than the MELD data set and it is therefore more appropriate for finetuning BERT.

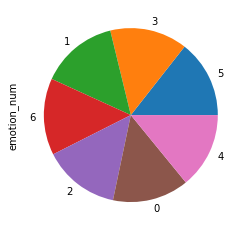
**The unique emotions, as annotated: 'joy', 'fear', 'anger', 'sadness', 'disgust', 'shame', 'guilt'**

There is no “neutral” category, which is a flaw of this data set. The emotions shame and guilt are not in the set we use for emotion detection from vision. This may still not be the ideal data set.

**The data set is balanced**

Field1 emotion\_num

1. 0 joy 0
2. 1 fear 1
3. 2 anger 2
4. 3 sadness 3
5. 4 disgust 4
6. 5 shame 5
7. 6 guilt 6

****

This is a very positive characteristic of this data set as, ideally, we would want to have the same learning opportunity for every emotion present.

For tokenization BertTokenizer was used from bert-base-uncased

After tokenization the train/test split was made with 80% training, 20% testing, random state 500.

Batch size was set to 16 (BERT creators recommend a batch size of 16 or 32 for finetuning BERT on a specific task)

The optimizer that was chosen was AdamW with learning rate 2e-5 and epsilon value 1e-8

Number of training epochs was set to 5

Below are the figures for the five epochs (accuracy, precision, recall and F1-score)

======== Epoch 1 / 5 ========

Training...

Batch 40 of 376. Elapsed: 0:00:14.

Batch 80 of 376. Elapsed: 0:00:27.

Batch 120 of 376. Elapsed: 0:00:41.

Batch 160 of 376. Elapsed: 0:00:55.

Batch 200 of 376. Elapsed: 0:01:09.

Batch 240 of 376. Elapsed: 0:01:22.

Batch 280 of 376. Elapsed: 0:01:36.

Batch 320 of 376. Elapsed: 0:01:50.

Batch 360 of 376. Elapsed: 0:02:03.

Average training loss: 0.61

Training epoch took: 0:02:09

Running Validation...

Accuracy: 0.70

Validation took: 0:00:10

precision recall f1-score support

0 0.8816327 0.9350649 0.9075630 231

1 0.7478632 0.8101852 0.7777778 216

2 0.6256410 0.5922330 0.6084788 206

3 0.6764706 0.7863248 0.7272727 234

4 0.7333333 0.6302083 0.6778711 192

5 0.5561497 0.5073171 0.5306122 205

6 0.6600985 0.6175115 0.6380952 217

accuracy 0.7035310 1501

macro avg 0.6973127 0.6969778 0.6953816 1501

weighted avg 0.6998160 0.7035310 0.6999127 1501

======== Epoch 2 / 5 ========

Training...

Batch 40 of 376. Elapsed: 0:00:14.

Batch 80 of 376. Elapsed: 0:00:27.

Batch 120 of 376. Elapsed: 0:00:41.

Batch 160 of 376. Elapsed: 0:00:55.

Batch 200 of 376. Elapsed: 0:01:09.

Batch 240 of 376. Elapsed: 0:01:22.

Batch 280 of 376. Elapsed: 0:01:36.

Batch 320 of 376. Elapsed: 0:01:50.

Batch 360 of 376. Elapsed: 0:02:03.

Average training loss: 0.50

Training epoch took: 0:02:09

Running Validation...

Accuracy: 0.71

Validation took: 0:00:10

precision recall f1-score support

0 0.8983051 0.9177489 0.9079229 231

1 0.8246445 0.8055556 0.8149883 216

2 0.5974026 0.6699029 0.6315789 206

3 0.7435897 0.7435897 0.7435897 234

4 0.6551724 0.6927083 0.6734177 192

5 0.5422886 0.5317073 0.5369458 205

6 0.6756757 0.5760369 0.6218905 217

accuracy 0.7095270 1501

macro avg 0.7052969 0.7053214 0.7043334 1501

weighted avg 0.7103801 0.7095270 0.7089895 1501

======== Epoch 3 / 5 ========

Training...

Batch 40 of 376. Elapsed: 0:00:14.

Batch 80 of 376. Elapsed: 0:00:27.

Batch 120 of 376. Elapsed: 0:00:41.

Batch 160 of 376. Elapsed: 0:00:55.

Batch 200 of 376. Elapsed: 0:01:09.

Batch 240 of 376. Elapsed: 0:01:22.

Batch 280 of 376. Elapsed: 0:01:36.

Batch 320 of 376. Elapsed: 0:01:50.

Batch 360 of 376. Elapsed: 0:02:03.

Average training loss: 0.33

Training epoch took: 0:02:09

Running Validation...

Accuracy: 0.71

Validation took: 0:00:10

precision recall f1-score support

0 0.9327354 0.9004329 0.9162996 231

1 0.7777778 0.8101852 0.7936508 216

2 0.5990991 0.6456311 0.6214953 206

3 0.7631579 0.7435897 0.7532468 234

4 0.6702128 0.6562500 0.6631579 192

5 0.5459184 0.5219512 0.5336658 205

6 0.6301370 0.6359447 0.6330275 217

accuracy 0.7068621 1501

macro avg 0.7027198 0.7019978 0.7020777 1501

weighted avg 0.7080540 0.7068621 0.7071794 1501

======== Epoch 4 / 5 ========

Training...

Batch 40 of 376. Elapsed: 0:00:14.

Batch 80 of 376. Elapsed: 0:00:27.

Batch 120 of 376. Elapsed: 0:00:41.

Batch 160 of 376. Elapsed: 0:00:55.

Batch 200 of 376. Elapsed: 0:01:09.

Batch 240 of 376. Elapsed: 0:01:22.

Batch 280 of 376. Elapsed: 0:01:36.

Batch 320 of 376. Elapsed: 0:01:50.

Batch 360 of 376. Elapsed: 0:02:03.

Average training loss: 0.25

Training epoch took: 0:02:09

Running Validation...

Accuracy: 0.71

Validation took: 0:00:10

precision recall f1-score support

0 0.9145299 0.9264069 0.9204301 231

1 0.7688889 0.8009259 0.7845805 216

2 0.6121495 0.6359223 0.6238095 206

3 0.7574468 0.7606838 0.7590618 234

4 0.6718750 0.6718750 0.6718750 192

5 0.5555556 0.5365854 0.5459057 205

6 0.6502463 0.6082949 0.6285714 217

accuracy 0.7108594 1501

macro avg 0.7043846 0.7058135 0.7048906 1501

weighted avg 0.7093099 0.7108594 0.7098765 1501

======== Epoch 5 / 5 ========

Training...

Batch 40 of 376. Elapsed: 0:00:14.

Batch 80 of 376. Elapsed: 0:00:27.

Batch 120 of 376. Elapsed: 0:00:41.

Batch 160 of 376. Elapsed: 0:00:55.

Batch 200 of 376. Elapsed: 0:01:09.

Batch 240 of 376. Elapsed: 0:01:22.

Batch 280 of 376. Elapsed: 0:01:36.

Batch 320 of 376. Elapsed: 0:01:50.

Batch 360 of 376. Elapsed: 0:02:03.

Average training loss: 0.21

Training epcoh took: 0:02:09

Running Validation...

Accuracy: 0.71

Validation took: 0:00:10

precision recall f1-score support

0 0.9145299 0.9264069 0.9204301 231

1 0.7688889 0.8009259 0.7845805 216

2 0.6121495 0.6359223 0.6238095 206

3 0.7574468 0.7606838 0.7590618 234

4 0.6718750 0.6718750 0.6718750 192

5 0.5555556 0.5365854 0.5459057 205

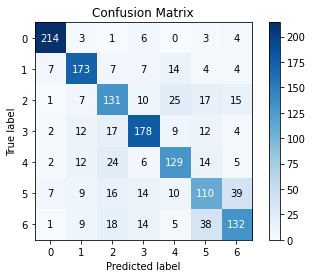
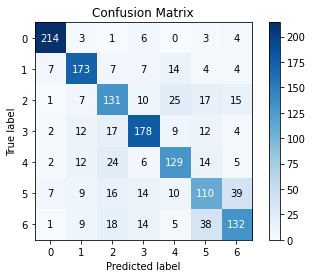
6 0.6502463 0.6082949 0.6285714 217

accuracy 0.7108594 1501

macro avg 0.7043846 0.7058135 0.7048906 1501

weighted avg 0.7093099 0.7108594 0.7098765 1501

The performance of the identical model jumps immediately with 10 percentage points to 71%. This is strictly due to the use of another data set. One which is populated much more densely and that is balanced.

Field1 emotion\_num

0 joy 0

1 fear 1

2 anger 2

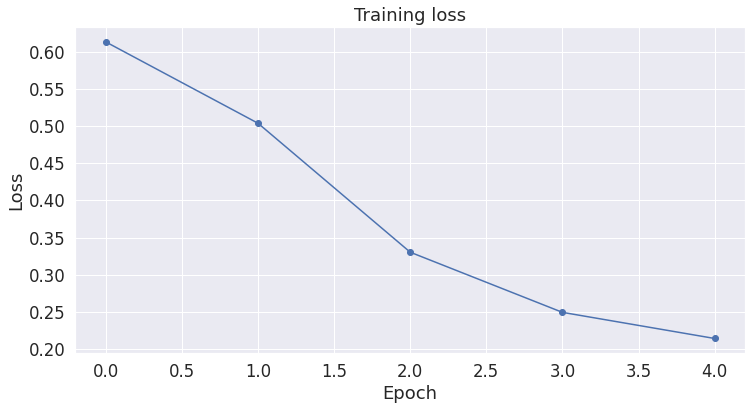
3 sadness 3

4 disgust 4

5 shame 5

6 guilt 6

The confusion matrices of epochs 4 and 5 respectively (the best ones with 71% accuracy score) show that shame and guilt are most often mistaken for each other. With disgust and anger in second place.



The training loss curve is not as linear as for the MELD data set.

**Conclusion**

Without tweaking anything, the ISEAR data set immediately improves the model’s results.

Below a comparison of the fine-tuned BERT model to a performance table published in March 2020.

Table 3 Summary of related work in emotion recognition in text from: [A survey of state-of-the-art approaches for emotion recognition in text](https://link.springer.com/article/10.1007/s10115-020-01449-0)[[18]](#footnote-18)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Reference | Method | Dataset | Measure | % |
| Ho and Cao | Learning-based | ISEAR | F-score | 35.3 |
| Kim et al. | Learning-based | ISEAR | F-score | 37.2 |
| Bandhakavi et al. | Learning-based | ISEAR | F-score | 39.48 |
| Udochukwu and He | Rule-based | ISEAR | F-score | 51.3 |
| Li et al. | Hybrid | ISEAR | F-score | 51.75 |
| Anusha and Sandhya | Learning-based | ISEAR | F-score | 63.1 |
| Herzig et al. | Hybrid | ISEAR | F-score | 64.1 |
| Riahi and Safari | Hybrid | ISEAR | F-score | 68 |
| Veller | BERT-based | ISEAR | F-score | 70 |
| Alotaibi, 2019\*[[19]](#footnote-19) | Machine Learning | ISEAR | F-score | 85 |
| Reference | Method | Dataset | Measure | % |
| Veller | BERT-based | ISEAR | Accuracy | 71 |
| Singh et al. | Learning-based | ISEAR | Accuracy | 72.43 |
| Thomas et al. | Learning-based | ISEAR | Accuracy | 76.96 |

1. **BERT model finetuned with a classifier on bespoke data set created by Lukas Garbas**

The dataset was combined from DailyDialog, ISEAR, and Emotion-Stimulus to create a balanced dataset with 5 labels: joy, sad, anger, fear, and neutral. The texts mainly consist of short messages and dialog utterances[[20]](#footnote-20).

**Number of sentences: 7933 (train) + 3393 (test)**

The total number of sentences is 11 326. Whereas the runs with the other data sets made a randomized and shuffled split between training and testing data of the full data set, this is not the case here, as we wanted to keep the training and testing data identical to be able to compare the models of Daniela and myself.

**The unique emotions, as annotated: ‘sadness’, 'neutral', 'anger', 'fear', 'joy'**

This data set has a “neutral” category. All the categories are in place to be able to concatenate the NLP model with Reginald’s vision model.

**The data set is balanced**

Emotion emotion\_num

sadness 0

neutral 1

anger 2

fear 3

joy 4

****

**The baseline model (BERT base)**

For tokenization BertTokenizer was used from bert-base-uncased. The choice for uncased was induced by the fact that the inputs would be from a STT system, where no casing is used to potentially express emotion, as is sometimes done in written text.

The two sets (training and validation) were concatenated into one big meta set for tokenization. After tokenization, the meta set was split again into the original training and testing set.

Batch size was set to 16 (BERT creators recommend a batch size of 16 or 32 for finetuning BERT on a specific task). Literature indicates that 16 is the optimal batch size for accuracy. Bigger batch sizes will lead to reduced accuracy.

The optimizer that was chosen was AdamW with learning rate 2e-5 and epsilon value 1e-8.

Number of training epochs was set to 5 as this was the maximum number of epochs specified in class and also in literature it is not recommended to increase the number of epochs beyond 5.

All parameters of the baseline model were kept the same as for the runs with the two other data sets, in order to be able to compare.

**Below are the figures for the five epochs (accuracy, precision, recall and F1-score)**

======== Epoch 1 / 5 ========

Training...

Batch 40 of 496. Elapsed: 0:01:22.

Batch 80 of 496. Elapsed: 0:02:45.

Batch 120 of 496. Elapsed: 0:04:09.

Batch 160 of 496. Elapsed: 0:05:32.

Batch 200 of 496. Elapsed: 0:06:55.

Batch 240 of 496. Elapsed: 0:08:19.

Batch 280 of 496. Elapsed: 0:09:42.

Batch 320 of 496. Elapsed: 0:11:05.

Batch 360 of 496. Elapsed: 0:12:28.

Batch 400 of 496. Elapsed: 0:13:51.

Batch 440 of 496. Elapsed: 0:15:15.

Batch 480 of 496. Elapsed: 0:16:38.

Average training loss: 0.83

Training epoch took: 0:17:11

Accuracy: 0.69

precision recall f1-score support

0 0.65372 0.70793 0.67974 1640

1 0.77559 0.73144 0.75287 1616

2 0.67727 0.65262 0.66472 1566

3 0.67115 0.66890 0.67002 1492

4 0.69212 0.69981 0.69595 1619

accuracy 0.69280 7933

macro avg 0.69397 0.69214 0.69266 7933

weighted avg 0.69431 0.69280 0.69315 7933

Running Validation...

Accuracy: 0.81

Validation took: 0:02:00

precision recall f1-score support

0 0.82671 0.76923 0.79693 676

1 0.82770 0.80564 0.81652 638

2 0.77240 0.78355 0.77794 693

3 0.81664 0.85272 0.83429 679

4 0.82490 0.85290 0.83866 707

accuracy 0.81314 3393

macro avg 0.81367 0.81281 0.81287 3393

weighted avg 0.81341 0.81314 0.81291 3393

======== Epoch 2 / 5 ========

Training...

Batch 40 of 496. Elapsed: 0:01:23.

Batch 80 of 496. Elapsed: 0:02:46.

Batch 120 of 496. Elapsed: 0:04:10.

Batch 160 of 496. Elapsed: 0:05:33.

Batch 200 of 496. Elapsed: 0:06:56.

Batch 240 of 496. Elapsed: 0:08:19.

Batch 280 of 496. Elapsed: 0:09:42.

Batch 320 of 496. Elapsed: 0:11:06.

Batch 360 of 496. Elapsed: 0:12:29.

Batch 400 of 496. Elapsed: 0:13:52.

Batch 440 of 496. Elapsed: 0:15:15.

Batch 480 of 496. Elapsed: 0:16:39.

Average training loss: 0.40

Training epoch took: 0:17:12

Accuracy: 0.87

precision recall f1-score support

0 0.88206 0.85732 0.86951 1640

1 0.86311 0.87005 0.86656 1616

2 0.84720 0.87803 0.86234 1566

3 0.90097 0.87198 0.88624 1492

4 0.87766 0.89067 0.88412 1619

accuracy 0.87357 7933

macro avg 0.87420 0.87361 0.87375 7933

weighted avg 0.87398 0.87357 0.87362 7933

Running Validation...

Accuracy: 0.83

Validation took: 0:02:00

precision recall f1-score support

0 0.78423 0.83876 0.81058 676

1 0.81623 0.83542 0.82572 638

2 0.78348 0.79365 0.78853 693

3 0.88480 0.81443 0.84816 679

4 0.86232 0.84158 0.85183 707

accuracy 0.82464 3393

macro avg 0.82621 0.82477 0.82496 3393

weighted avg 0.82649 0.82464 0.82504 3393

======== Epoch 3 / 5 ========

Training...

Batch 40 of 496. Elapsed: 0:01:23.

Batch 80 of 496. Elapsed: 0:02:46.

Batch 120 of 496. Elapsed: 0:04:09.

Batch 160 of 496. Elapsed: 0:05:33.

Batch 200 of 496. Elapsed: 0:06:56.

Batch 240 of 496. Elapsed: 0:08:19.

Batch 280 of 496. Elapsed: 0:09:42.

Batch 320 of 496. Elapsed: 0:11:05.

Batch 360 of 496. Elapsed: 0:12:29.

Batch 400 of 496. Elapsed: 0:13:52.

Batch 440 of 496. Elapsed: 0:15:15.

Batch 480 of 496. Elapsed: 0:16:38.

Average training loss: 0.23

Training epoch took: 0:17:11

Accuracy: 0.93

precision recall f1-score support

0 0.93097 0.92927 0.93012 1640

1 0.91278 0.91955 0.91615 1616

2 0.93333 0.93870 0.93601 1566

3 0.94952 0.93298 0.94118 1492

4 0.93485 0.93947 0.93715 1619

accuracy 0.93193 7933

macro avg 0.93229 0.93199 0.93212 7933

weighted avg 0.93201 0.93193 0.93195 7933

Running Validation...

Accuracy: 0.82

Validation took: 0:02:00

precision recall f1-score support

0 0.80119 0.79882 0.80000 676

1 0.81746 0.80721 0.81230 638

2 0.74935 0.83694 0.79073 693

3 0.87809 0.83800 0.85757 679

4 0.86807 0.81895 0.84279 707

accuracy 0.82022 3393

macro avg 0.82283 0.81998 0.82068 3393

weighted avg 0.82298 0.82022 0.82086 3393

======== Epoch 4 / 5 ========

Training...

Batch 40 of 496. Elapsed: 0:01:23.

Batch 80 of 496. Elapsed: 0:02:46.

Batch 120 of 496. Elapsed: 0:04:10.

Batch 160 of 496. Elapsed: 0:05:33.

Batch 200 of 496. Elapsed: 0:06:56.

Batch 240 of 496. Elapsed: 0:08:20.

Batch 280 of 496. Elapsed: 0:09:43.

Batch 320 of 496. Elapsed: 0:11:06.

Batch 360 of 496. Elapsed: 0:12:29.

Batch 400 of 496. Elapsed: 0:13:53.

Batch 440 of 496. Elapsed: 0:15:16.

Batch 480 of 496. Elapsed: 0:16:39.

Average training loss: 0.14

Training epoch took: 0:17:12

Accuracy: 0.96

precision recall f1-score support

0 0.96267 0.95915 0.96090 1640

1 0.94370 0.95421 0.94892 1616

2 0.96574 0.97190 0.96881 1566

3 0.97823 0.96381 0.97097 1492

4 0.96170 0.96170 0.96170 1619

accuracy 0.96206 7933

macro avg 0.96241 0.96215 0.96226 7933

weighted avg 0.96214 0.96206 0.96208 7933

Running Validation...

Accuracy: 0.82

Validation took: 0:02:00

precision recall f1-score support

0 0.81778 0.81657 0.81717 676

1 0.80211 0.83229 0.81692 638

2 0.79094 0.78066 0.78577 693

3 0.85952 0.83800 0.84862 679

4 0.84789 0.85149 0.84968 707

accuracy 0.82375 3393

macro avg 0.82365 0.82380 0.82363 3393

weighted avg 0.82398 0.82375 0.82378 3393

======== Epoch 5 / 5 ========

Training...

Batch 40 of 496. Elapsed: 0:01:23.

Batch 80 of 496. Elapsed: 0:02:46.

Batch 120 of 496. Elapsed: 0:04:09.

Batch 160 of 496. Elapsed: 0:05:33.

Batch 200 of 496. Elapsed: 0:06:56.

Batch 240 of 496. Elapsed: 0:08:19.

Batch 280 of 496. Elapsed: 0:09:42.

Batch 320 of 496. Elapsed: 0:11:05.

Batch 360 of 496. Elapsed: 0:12:28.

Batch 400 of 496. Elapsed: 0:13:51.

Batch 440 of 496. Elapsed: 0:15:15.

Batch 480 of 496. Elapsed: 0:16:38.

Average training loss: 0.10

Training epoch took: 0:17:11

Accuracy: 0.98

precision recall f1-score support

0 0.97672 0.97195 0.97433 1640

1 0.95971 0.97277 0.96620 1616

2 0.97893 0.97893 0.97893 1566

3 0.98518 0.97989 0.98253 1492

4 0.97768 0.97406 0.97587 1619

accuracy 0.97542 7933

macro avg 0.97564 0.97552 0.97557 7933

weighted avg 0.97548 0.97542 0.97544 7933

Running Validation...

Accuracy: 0.82

Validation took: 0:02:00

precision recall f1-score support

0 0.80584 0.81657 0.81117 676

1 0.79468 0.84326 0.81825 638

2 0.80149 0.77489 0.78797 693

3 0.85952 0.83800 0.84862 679

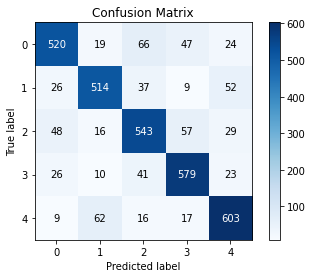
4 0.85265 0.84300 0.84780 707

accuracy 0.82287 3393

macro avg 0.82284 0.82314 0.82276 3393

weighted avg 0.82335 0.82287 0.82289 3393

The performance of the identical (untweaked) model that was used with the MELD and ISEAR jumps again with 10 percentage points to 83% for the best runs. We can now confirm that the chosen data set is quintessential for the performance of the model. The more utterances the better the performance and the balance of the data set is also key.



Emotion emotion\_num

sadness 0

neutral 1

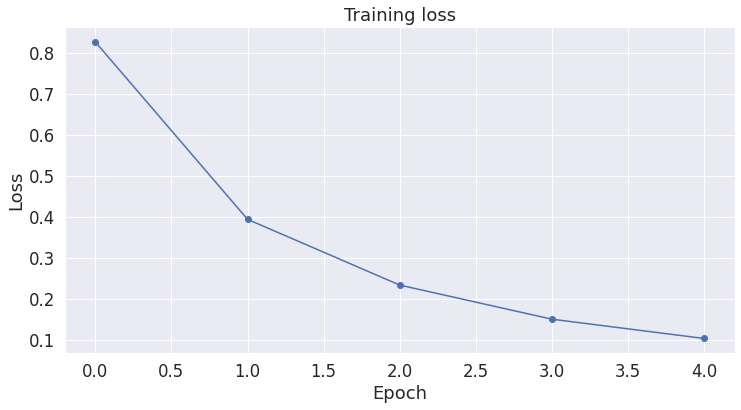
anger 2

fear 3

joy 4

The confusion matrices of epoch 2 (best run) (0.82464 accuracy score) shows that anger and sadness are most often mistaken for each other. With joy and neutral in second place.

It is clear from the accuracy figures for the training and testing that the model starts overfitting after 2 epochs.



The training loss curve is gradual.

**Conclusion**

Without tweaking anything, the Lukas Garbas data set improves the model’s results even more compared to the performance with the ISEAR data set (again a jump of +10% Accuracy score to 83%).

Because this is a bespoke data set, we cannot make benchmark comparisons. We can however compare to the results that Lukas Garbas got on his data set, which obtained an F1-score of 0.832. Our best model meets that benchmark. In a next step our model will be run with certain parameter tweaks, to see if the performance can be improved.

1. **Polishing and tweaking the BERT-based model with the Lukas Garbas data set**

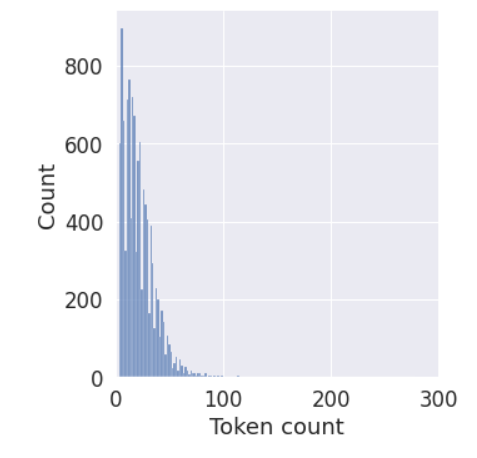
The table with the different tweaks (39) can be found in Annex 2.

The following parameters were used/changed:

**BERT model:** The Bert base pretrained model was used for all the runs. BERT large could not be used because of insufficient access to an appropriate GPU to use Bert large (BERT large has 24 layers instead of BERT base’s 12 layers).

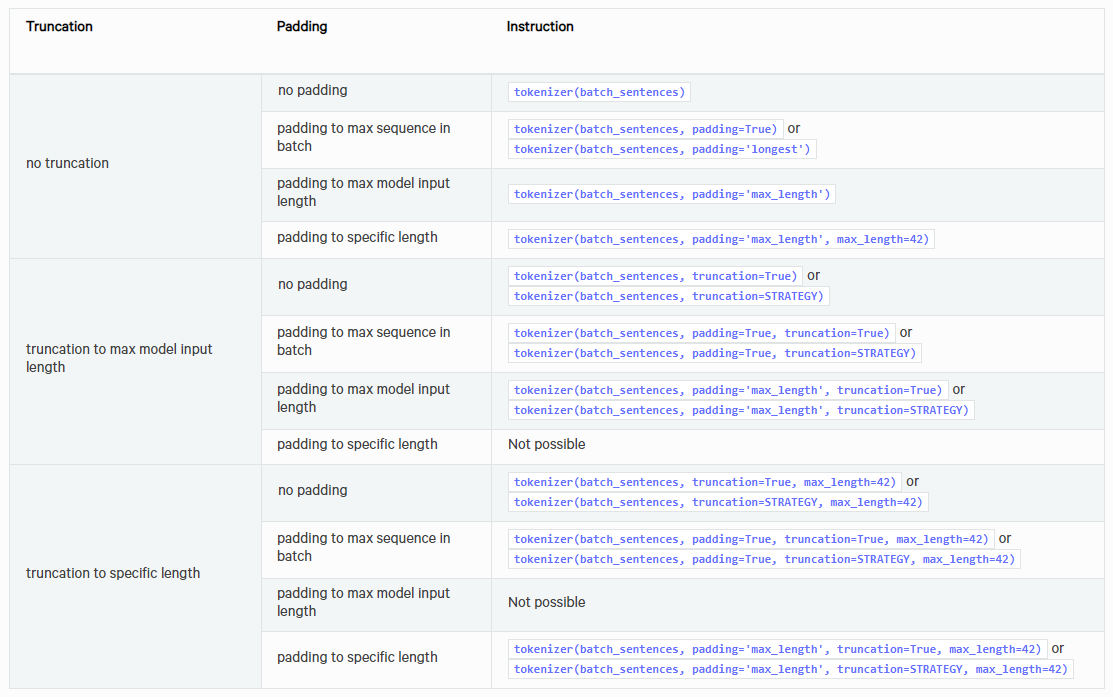
**Tokenizer**: It was not possible to use a different tokenizer as BERT only works with the BERT tokenizer that “translates” each token to an ID in the BERT vocabulary. The model used bert-base-uncased.

**Padding**: it was interesting to see if the padding makes any difference to the accuracy. Initially the padding was set to the maximum length of the utterances plus a few more as buffer (300). A histogram however shows that there are only a few very long utterances.



Padding can therefore be reduced. In the experiment the padding was reduced in individual runs to 100, 150, 200 and 277 respectively. Reducing the padding did not have a significant effect on the performance of the model. In hindsight, it is unclear if it is possible to reduce the padding in this simple way, as the model tokenizes the sentences first, thereby adding the special tokens [CLS] at the start and [SEP] at the end of the utterance. These two tokens are obligatory. It is not clear what happens to the [SEP] token if we reduce the padding to before the end of the sentence (and therefore before the [SEP] token). It is possible that these sentences are simply left out (as they lack the required terminator token). This is plausible, as the model does not throw any errors when we do the simple padding reduction. This could however mean that we are creating an imbalance in the data set if all the longer sentences have the same emotion tag for instance. The literature did not make us any wiser, therefore the only thing we can conclude is that reducing the padding in this simplistic way does not reliably improve the accuracy of the model. The uncertainty was fixed by looking at the ID of the token before the first [PAD] token and changing it to 202 (the ID of the [SEP] token) if it was not 202.

There are more sophisticated ways to alter the padding, such as sorting in batches according to length of the utterances and then setting MAX\_LEN to the length of the longest utterance in the batch (per batch), but this becomes rather complex and it was more fruitful to try to change other parameters first.

[[21]](#footnote-21)

It should be noted that reducing the length of the utterances does allow us to run bigger batches (up to batch size 64) without there necessarily being a penalty on the Accuracy.

**Shuffle data**: The data were not shuffled in the training/validation split, in order to keep the training and validation sets intact.

**Batch size**: Recommended batch sizes are 16 and 32 for BERT classifier models. The baseline model had the batch size set to 12. The literature mostly reports that increasing the batch size will improve the training speed, but it will also reduce Accuracy. At first it was the intention to keep the batch size to 16, as the aim is to achieve maximum Accuracy, but when the batch size was set to 32 it did not necessarily affect Accuracy. This was very much dependent upon other parameter values. In the end, the best performing model had the batch size set to 32.

**Data loader**: The data loader remained unchanged from the baseline model, with a RandomSampler for the training data and a SequentialSampler for the validation data. The RandomSampler was chosen for the training data to introduce more randomness in the sequence to help avoiding overfitting.

**Optimizer**: The optimizer was kept unchanged (AdamW) because earlier experiments with other optimizers on the smaller data sets had proven unfruitful. Also, from literature it appears that the AdamW optimizer can be seen as a state-of the-art optimizer as already explained for the baseline model. The AdamW optimizer decouples the weight decay from the optimization step. This means that the weight decay and learning rate can be optimized separately, i.e. changing the learning rate does not change the optimal weight decay. According to the literature this allows for improved generalization performance. According to that same literature on AdamW warm restarts could be used (they could not be used on the AdamW predecessor Adam, due to its dysfunctional weight decay) In each restart, the learning rate is initialized to some value and is scheduled to decrease. Importantly, the restart is warm as the optimization does not start from scratch but from the parameters to which the model converged during the last step. The key factor is that the learning rate is decreased with an aggressive cosine annealing schedule, which rapidly lowers the learning rate[[22]](#footnote-22). As this is quite advanced material and this is but a beginner project the warm restart was not introduced into the model.

**Learning rate**: The learning rate was kept small as advised in class. The literature advises to set the learning rates for the AdamW optimizer to either 5e-5, 3e-5 or 2e-5. All the learning rates were tried in the experiment in different constellations. Two trials with higher learning rates ended at the bottom of the ranking, whereby the models just blatantly did not learn. Although it was clear in advance that learning rates should be set small, it was interesting to see what happens when the advice is ignored. The typical learning rate value for AdamW in the literature is 3,00E-04, but for our purposes smaller is clearly better. Our best model had a learning rate of 3,00E-05. The BERT pretrained’s AdamW default learning rate is 1,00E-04, which is why this is expressed as 0,0001 in the Excel table.

**Epsilon value**: The epsilon value was set to 1,00E-08 in the baseline model and it was not changed, as everywhere in literature this was the recommended value for AdamW.

**Epochs**: The original number of epochs of the baseline model was set to five, because this is the maximum number of recommended epochs for BERT. In the earlier experiments it had very rapidly become clear that the Accuracy results of the finetuning with the classifier started to diverge, usually from the second epoch onwards. For the Lukas Garbas data set the same effect was noticeable. While the training accuracy goes up as high as 91%, the training accuracy always remains in the low 80%. While looking at the results of the interim epochs something odd was noticed. Sometimes in a run of e.g. 4 epochs, the best Accuracy result would be reached at e.g. epoch 3. When the same model was re-run with exactly the same parameters, but with only 3 epochs, the final Accuracy would be different from the previous run with 4 epochs. It is still unclear why this would happen, as the seed value was kept unchanged for all the runs and no randomization was introduced in splitting the training and valuation sets. It is however possible that the differences are due to the data loader, that does introduce some randomness to the training data with its RandomSampling method.

There is no clear line in the performance of the model with respect to more or less epochs between 2 and 4, results seem to be dependent on the number of warmup steps and the shape of scheduler, although 4 epochs seems to be over-represented in the top-10 of the experiment. This however could be due to overfitting. We already saw that training and testing Accuracy diverges, usually even after 2 epochs.

To avoid overfitting the early stopping method was introduced at the last stage of the experiment.

**The scheduler**: two schedulers were used in the experiment. A linear schedule with warmup and a cosine schedule with warmup. No conclusions can be drawn as to which one would be better, they seem to be equally adequate. The warmup was chosen as in the literature it is reported that this allows for better generalization of the model.

**Warmup steps**: No real conclusions can be drawn. The best model has 0 warmup steps, the second best 120. The effect can potentially only be seen on the test set, as the warmup is supposed to aid with generalization of the model.

The best result (without early stopping) was achieved with the parameters set as follows:

|  |  |  |
| --- | --- | --- |
|  | Best | Baseline |
| Model | bert-base-uncased | bert-base-uncased |
| Tokenizer | bert-base-uncased | bert-base-uncased |
| Padding | 300 | 300 |
| Shuffle data | FALSE | FALSE |
| Batch size | 32 | 16 |
| Data loader | Random sampler/Sequential Sampler | Random sampler/Sequential Sampler |
| Optimizer | AdamW | AdamW |
| Learning rate | 3,00E-5 | 3,00E-5 |
| Epsilon | 1,00E-08 | 1,00E-08 |
| Epochs | 4 | 2 |
| Scheduler | Linear schedule with warmup | Linear schedule with warmup |
| Warmup steps | 80 | 0 |
| Accuracy | 0,8311 | 0,8158 |

On the excel in the annex the base model with 0.82464 Accuracy score is highlighted in red. Note that this best Accuracy was noted in the second of four epochs in this run. While tweaking the model it has become clear that the results are not the same when only two epochs are programmed, even for otherwise identical parameters.

Note that the validation accuracy of 83% was already reached after the first epoch and that the full execution of all four epochs may mean that the model is overtrained. This is why early stopping should be introduced.

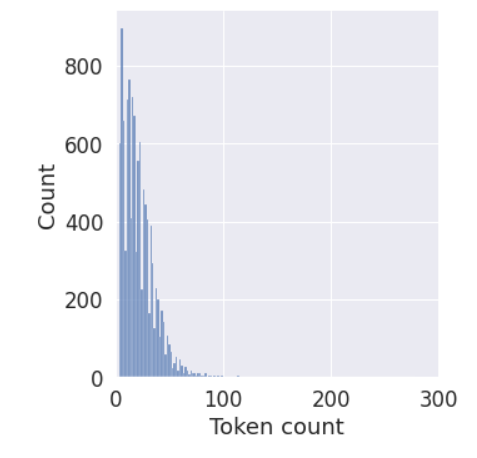
Limiting the padding has an effect, but does not seem to be necessary, in view of the fact that the best model had the original baseline padding of 300.

Contrary to what the literature claims, batch size of 32 (vs. 16) can increase the Accuracy and even runs with batch sizes of 64 may score better than the baseline model. There are too many parameter combinations to draw a conclusion with respect as to what has the most impact, which in a sense is unsatisfying because it is unclear as to why a model is performing better or worse.

The best result (with early stopping) was achieved with the parameters set as follows:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Best early stopping | Best | Baseline |
| Model | bert-base-uncased | bert-base-uncased | bert-base-uncased |
| Tokenizer | bert-base-uncased | bert-base-uncased | bert-base-uncased |
| Padding | 100 | 300 | 300 |
| Shuffle data | FALSE | FALSE | FALSE |
| Batch size | 32 | 32 | 16 |
| Data loader | Random sampler/Sequential Sampler | Random sampler/Sequential Sampler | Random sampler/Sequential Sampler |
| Optimizer | AdamW | AdamW | AdamW |
| Learning rate | 3,00E-5 | 3,00E-5 | 3,00E-5 |
| Epsilon | 1,00E-08 | 1,00E-08 | 1,00E-08 |
| Epochs | 4 | 4 | 2 |
| Scheduler | Cosine schedule with warmup | Linear schedule with warmup | Linear schedule with warmup |
| Warmup steps | 120 | 80 | 0 |
| Accuracy | 0,8302 | 0,8311 | 0,8158 |
| Stop at epoch | 2 | / | / |

The assumption is that padding to less than maximum utterance length has taken out parts of some utterances thereby potentially creating imbalances or inaccuracies in the data sets (by eliminating parts of sentences that may be significant for emotion determination) but the histogram shows that this effect may, overall, be negligible.



In view of the fact that for almost every run there was a divergence of training Accuracy and evaluation Accuracy after epoch 2, it can be concluded fairly reliably that the Accuracy figure for the early stopping run is closer to true Accuracy value than for the former record holder model. It should be noted that the early stopping model also meets (but doesn’t beat) the benchmark of 83% for the Lukas Garbas data set.

1. **Different split of the Lukas Garbas data set to be able to compare NLP Model 1 with NLP Model 2**

To allow comparison with the performance of NLP Model 2 the hyperparameters were used from the BERT model with the best statistics (accuracy, F1) and the same data set was used (Lukas Garbas), however, the Train data set was now split into a 70/30% Train and Validation set (no randomization, no shuffle to keep the data as conform as possible to those used for NLP Model 2). The Test set that was used as validation set in the above process was now used as Test set for the trained model.

The following results were noted:

precision recall f1-score support

0 0.81089 0.79290 0.80180 676

1 0.78752 0.83072 0.80854 638

2 0.79732 0.77201 0.78446 693

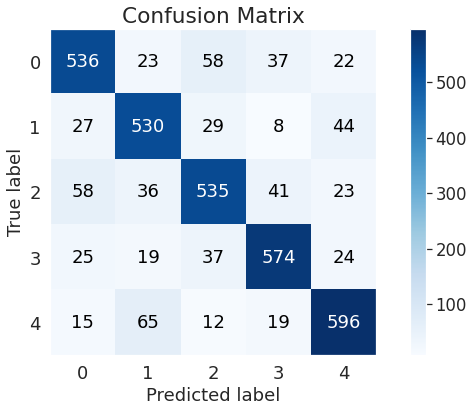
3 0.84536 0.84536 0.84536 679

4 0.84062 0.84300 0.84181 707

accuracy 0.81668 3393

macro avg 0.81634 0.81680 0.81639 3393

weighted avg 0.81682 0.81668 0.81658 3393

****

Emotion emotion\_num

sadness 0

neutral 1

anger 2

fear 3

joy 4

The same confusions can be detected but now with joy/neutral confusion in first place and anger/sadness in second. It seems that in analysing human language with this pretrained BERT model that these are the most frequent confusions. For our application a confusion between anger and sadness may not be such a good outcome, although both emotions are on the negative spectrum.

An **F1 score** of **81.6 %** on the Test set is quite respectable, i.e. the model performs quite well (statistically) on new data.

This is the model used for the Goldfish application for which the following code was supplied:

# Environmnet requirements

# pip3 install torch==1.8.1+cpu torchvision==0.9.1+cpu torchaudio===0.8.1 -f https://download.pytorch.org/whl/torch\_stable.html

# pip3 install transformers

# pip3 install pandas

import torch

import pandas as pd

from transformers import BertTokenizer, BertForSequenceClassification

import numpy as np

from keras\_preprocessing.sequence import pad\_sequences

from datetime import datetime

# Run on the CPU

device = torch.device("cpu")

# Load previously saved tokenizer

tokenizer\_path = './output\_trained\_model'

tokenizer = BertTokenizer.from\_pretrained(tokenizer\_path, do\_lower\_case=True)

# Load previously saved model

model\_path = './output\_trained\_model'

model = BertForSequenceClassification.from\_pretrained(model\_path)

# Function that applies the trained BERT model to inputed sentences

def validate\_emo(model, tokenizer, utterances, max\_lenght=0):

    # Tokenize the utterances

    input\_ids = []

    for utt in utterances:

        encoded\_sent = tokenizer.encode(

                            utt,                        # utterances to encode

                            add\_special\_tokens = True,  # Add '[CLS]' and '[SEP]'

                    )

        input\_ids.append(encoded\_sent)

    # pad the encoded input\_ids

    if max\_lenght == 0:

        max\_lenght = len(encoded\_sent)

    input\_ids = pad\_sequences(input\_ids, maxlen=max\_lenght, dtype="long",

                          value=0, truncating="post", padding="post")

    # Create attention masks

    attention\_masks = []

    # For each utterance...

    for sent in input\_ids:

    # Create the attention mask.

        #   - If token ID is 0, then it is padding => set mask to 0

        #   - If token ID is > 0, then it is a real token => set mask to 1

        att\_mask = [float(token\_id > 0) for token\_id in sent]

        # Store attention mask for utterance

        attention\_masks.append(att\_mask)

    # Convert all inputs into torch tensors

    validation\_inputs = torch.tensor(input\_ids)

    validation\_masks = torch.tensor(attention\_masks)

    model.eval()

    with torch.no\_grad():

        outputs = model(validation\_inputs,

                        token\_type\_ids=None,

                        attention\_mask=validation\_masks)

    logits = outputs[0]

    logits\_flat = np.argmax(logits, axis=1).flatten()

    emotion = logits\_flat[0].item()

    return emotion

# Turn tensor output into appropriate emotion str

def emotion\_int\_to\_str(emo\_int):

    if emo\_int == 0:

        return 'sadness'

    elif emo\_int == 1:

        return 'neutral'

    elif emo\_int == 2:

        return 'anger'

    elif emo\_int == 3:

        return 'fear'

    elif emo\_int == 4:

        return 'joy'

    else:

        return 'unknown'

# Input the sentence to be analysed for emotion

sentence = input("Type a sentence: ")

# Call the function that applies the trained BERT model

emotion = validate\_emo(model=model, tokenizer=tokenizer,utterances=[sentence],max\_lenght=100)

# Print datetime, emotion and tensor output (for verification purposes)

# Datetime and emotion are to be sent to text file or Excel for Doctor to assess

now = datetime.now()

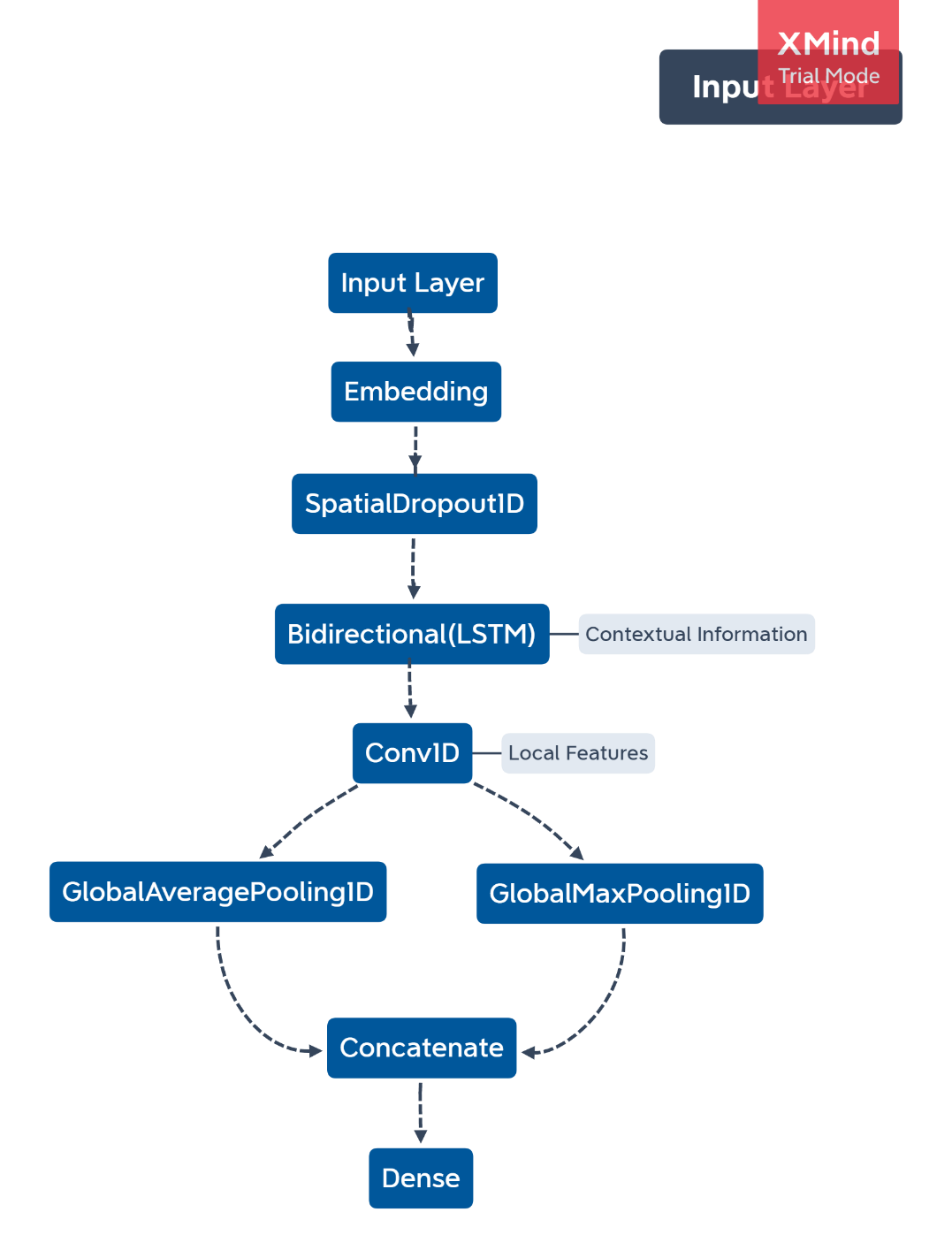
print(now)

print(emotion\_int\_to\_str(emotion))

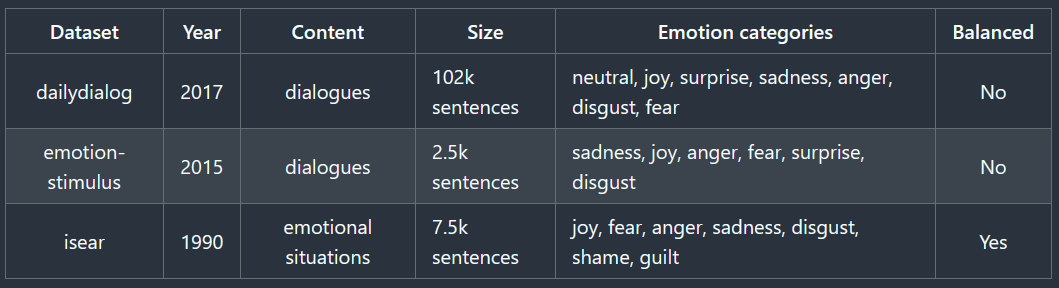
print(emotion)

Because of the discovery of the potential pre-training bias in very large models such as BERT, but also GPT3 and later models, and because it was suggested at the onset of our assignment that older models may provide equally good results, a second NLP model was developed.

1. **NLP Model 2:**

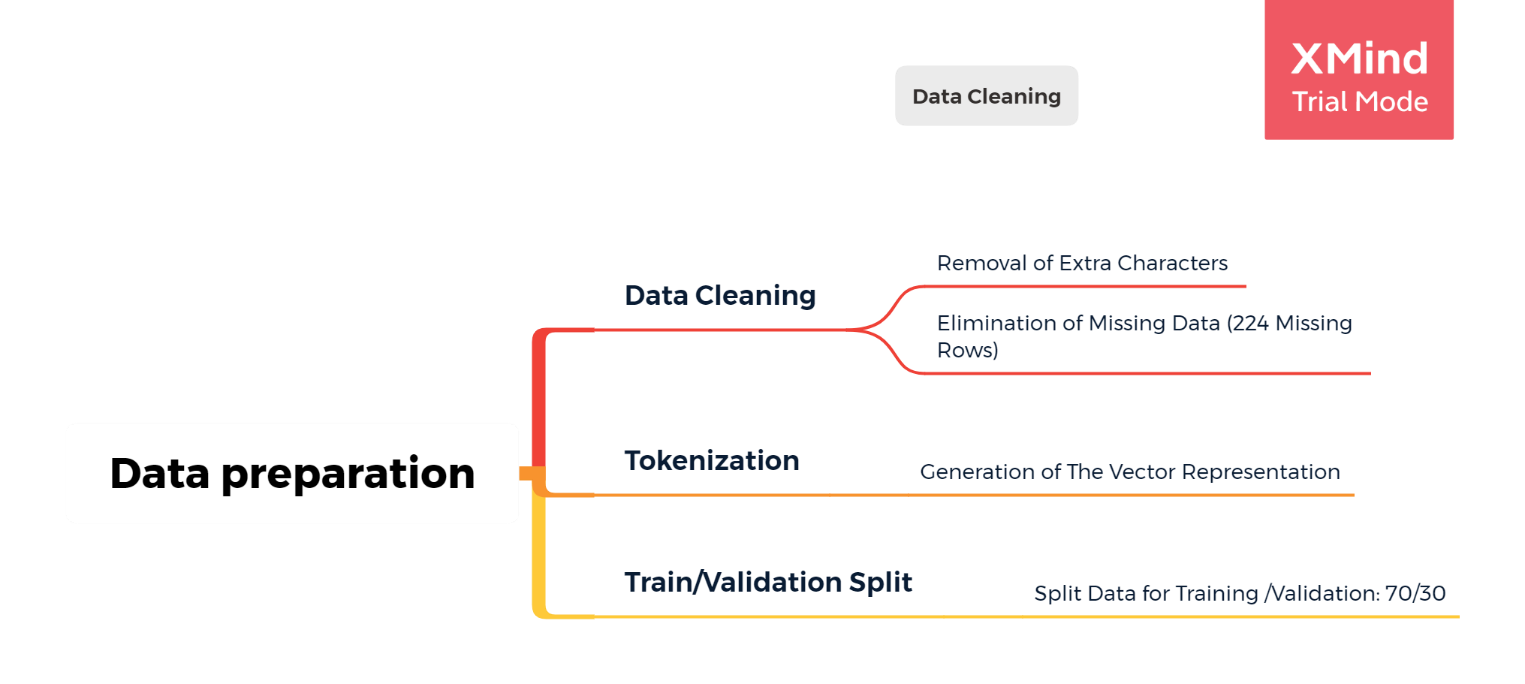
As previously mentioned, the NLP2 model comprises a Bidirectional LSTM (a type of Recurrent Neural Network) with a CNN (Convolutional Neural Network) layer. This graph was made to further clarify the architecture of the model.

*Figure x – NLP2 Model Architecture*

One of the first steps was tokenizing the data. The model was trained using the same dataset as the BERT model, the combined dataset by Lukas Garbas. As on this table below are listed the datasets that this dataset combines.

*Figure x – Datasets included in the employed dataset.*

The first step consisted of the cleaning of the data. The dataset was cleaned by the removal of extra characters from the data, we also removed missing data (in this case empty rows), which were in total 224 rows. Secondly, the NLP data was tokenized, transforming the text corpus into a vector representation. After this, the train/test split was done. In which we split data for training and for validation (70/30). It was not necessary to split into a test set as the original dataset was already split into two datasets – one for training and one for testing.



*Figure x – Mind-map of the preprocessing steps.*

The model was tuned through the change of hyperparameters. On the table below are presented some of the values/iterations explored during this tuning. Below are displayed the values for each of the model’s the training loss, training accuracy, validation loss and validation accuracy. For the training, different combinations of batch sizes and epochs were made, and the results were compared. One thing that was noticed was that as the epochs were increased, the model started overfitting the data, therefore, the number of epochs was kept at 2, which seems to be the optimal value.

Out of all the generated models, the last 3 presented on this table are the ones that had the highest performance. From those 3 the last one was chosen (model 18) as its performance is closer in the training and the validation sets. The last two models, even though they had a higher level of accuracy, had bigger gaps between the values for the training and validation sets, therefore they are more likely to be overfitting, so those were excluded.

Uma imagem com mesa

Descrição gerada automaticamente

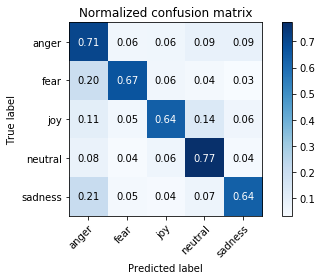
*Table x – Model Tuning*

There overall f-score was also computed, as a metric to compare the generated models with other models explored during the state of art exploration and the NLP1 BERT model. In the below table are presented the the achieved overall f-scores for the 3 best previously shown models.

Regarding the last model (m18), for this one the f-score was of 68.1%, which was inferior to BERT (81.6%) and other models we analyzed during the state of art exploration.

*Table x – F-scores for the 3 Best models*

Below is presented the normalized confusion matrix. The model’s best performance is in the prediction of “Neutral”, followed by anger with 0.71 (of the classifications being True Positives). The worse predictions are joy and sadness with only .64 true positives. Even though this model has a lower performance than the BERT model this might still be relevant model for the project particularly because of the issues surrounding BERT that were previously presented. This model might actually be learning more from the dataset, than the BERT model.



*Figure x – The normalized confusion matrix for model m18*

**Other text to emotion models:**

<https://www.youtube.com/watch?v=-CAC4wK9Ey0&list=PLEJK-H61XlwxpfpVzt3oDLQ8vr1XiEhev>

Text Classification | Sentiment Analysis with BERT using huggingface, PyTorch and Python Tutorial : <https://www.youtube.com/watch?v=8N-nM3QW7O0>

<https://onlinelibrary.wiley.com/doi/full/10.1002/eng2.12189>

<https://www.tensorflow.org/tutorials/text/classify_text_with_bert>

<https://pypi.org/project/text2emotion/?fbclid=IwAR0GWOPl1aVZJZWx0OzVTOc_lb5UZu_XRt27jOMf7dveb5OJShpsOFuwiEw>

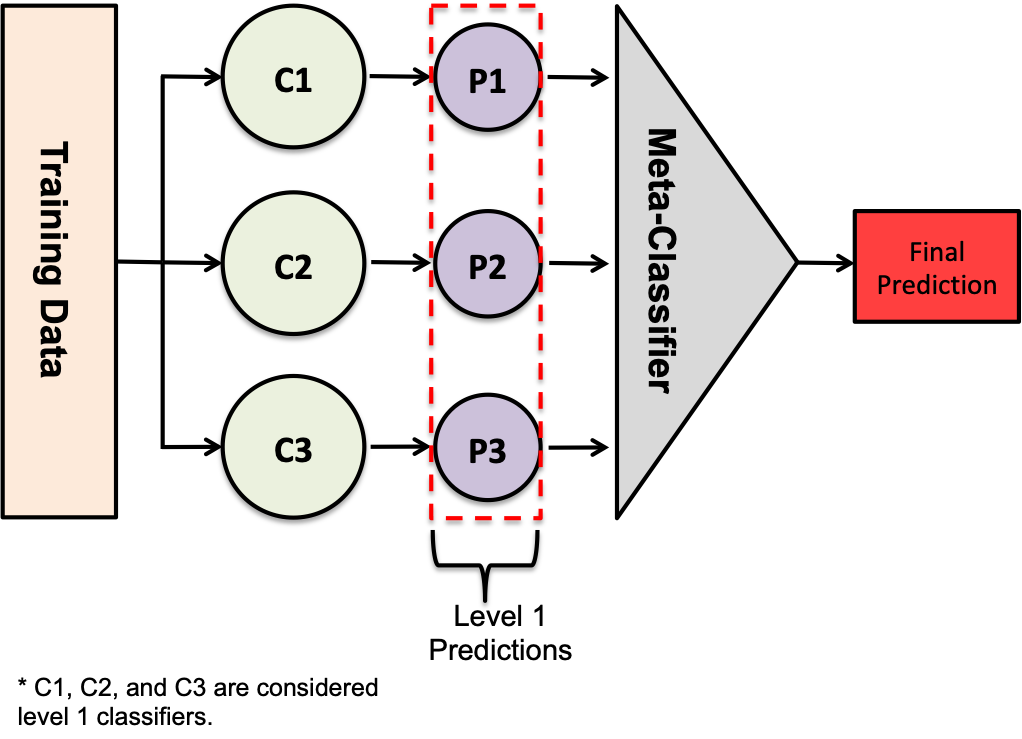
<https://colab.research.google.com/drive/1sCAcIGk2q9dL8dpFYddnsUin2MlhjaRw?usp=sharing#scrollTo=ivUkOaBPEQYr>

<https://www.tensorflow.org/tutorials/keras/text_classification>

# Concatenation of models

Because we started this project with the intention to create a multimodal model, but in the end, we could only obtain the tensor outputs of the two NLP models we decided to concatenate NLP1 and NLP2 to see if we could create a model that would outperform models 1 and 2, hereby creating a “proof of concept” for concatenating multimodal models.

At the same time, this was a test to see if we could further improve the performance of the two models by using an ensemble model combining the best NLP1 and NLP2 models. We chose to stack the models and feed them to a classifier neural network.



*Figure: 10 Schematic of ensemble classifier using stacking*

From both trained NLP models, we extracted the tensors from the last hidden layer for the entire Train set of the Lukas Garbas data set.

We then merged the two data sets (the tensors) with Pandas dataframes and the merged data set was fed to a pre-existing neural network for which the hyperparameters were adjusted for the fact that we were working with already trained systems. The model did not want to learn if we did not adjust these hyperparameters. The code was adapted from a multilabel classifier for the Kaggle Wine challenge ( <https://towardsdatascience.com/pytorch-tabular-multiclass-classification-9f8211a123ab> ).

We did not have to modify the code much, we adjusted the number of Epochs to 100 (from 300) to prevent overtraining and the learning rate was reduced from 0.0007 to 0.00001 in order to allow the system to still learn. This was the most significant change we had to apply. The number of classes was set to 5 in view of the number of emotions (the wine example had 6 classes).

EPOCHS = 100

BATCH\_SIZE = 16

LEARNING\_RATE = 0.00001

NUM\_FEATURES = len(X.columns)

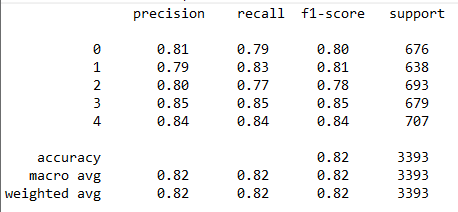
NUM\_CLASSES = 5

This model again produces prediction tensors whereupon an argmax layer is added to produce the final emotion output. We call this model NLP3.

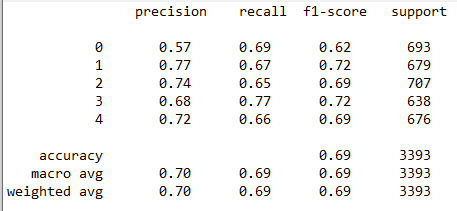
*Results:*

These were the scores for the best performing individual models as trained with the full Train set and scored against the same Test set.

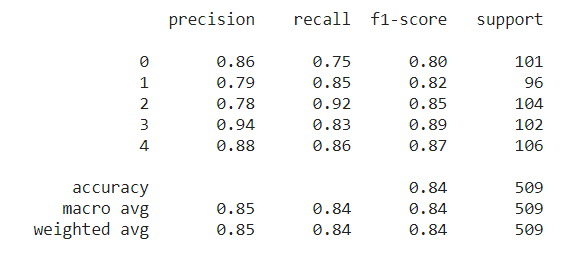
NLP1



NLP2



The results of the stacked model were as follows:



We can see that the stacked model scores an F1 score of 84% which is higher than for both NLP1 and NLP2.

We had to make adjustments to the Test data set for the stacked model, because we could not use the Train data set we had used for training NLP1 and NLP2. This gave worse results than those for our best model (NLP1). When we tried to figure out why this would be the case, we read that for ensemble models we should not use the Train data set that the level 1 classifiers were trained on.

We therefore took the Test data set (from the Lukas Garbas data sets) and divided this up in 55% Train data, 20% Validation data and 15% Test data. It is with these data that we achieved the result above.

*Conclusion*

The performance of NLP3 is better. With 84% it is above the benchmark set by Lukas Garbas. This is an indication that combining models indeed provides better results and we could extrapolate from this that it would be a good way to train multimodal models.

However, this last part was not self-evident for us to implement, and we do not know whether our methodology is truly correct or not. We should also add that this methodology puts the models in a pipeline, where the results of the level 1 models are fed to the next model, thereby propagating all the biases and potential errors of the level 1 models.

We have previously stated that we have concerns with respect to the BERT-based NLP1 model, and in view of the fact that the grounds for our concern propagate down the pipeline, we have decided not to implement NLP3 in the Goldfish and to keep the outputs of both NLP1 and NLP2 in the reporting to the medical care giver.

We are not convinced that any of these models, how high their F1 scores may be, truly represent human emotion. We think it cannot currently be done even with state-of-the art AI technology and we therefore conclude that our enthusiasm from the beginning of the project to make an application that aspires to do emotion detection and recognition was premature. It was of course very useful to do the exercise if only to gain the above insight and to understand the technology better.

# System 2 - Goldfish Backchanneling – data

The cartoon goldfish must be able to backchannel in real time as the person speaks.

In view of the fact that the multimodal model is too slow to be used in this fashion, we have chosen to separate the multimodal emotion detection from the backchanneling mechanism. The emotion detection for backchanneling purposes is done by a rule-based method with reference lexicon, inspired by the chatbot Eliza.

*9.1. Emotions data set*s

The following were the two lexica that fitted our requirements best:

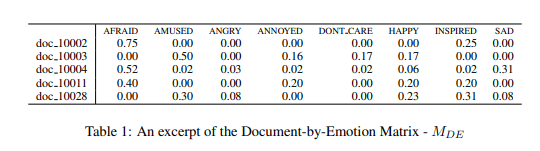
*NRC Word-Emotion Association Lexicon (EmoLex)*

This is a list of English words and their associations with eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive). The annotations were manually done by crowdsourcing.

<https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm#:~:text=NRC%20Word%2DEmotion%20Association%20Lexicon,were%20manually%20done%20by%20crowdsourcing>.

*Depeche Mood: a Lexicon for Emotion Analysis from Crowd Annotated News*

<https://www.aclweb.org/anthology/P14-2070>/



*Figure 11 - ?*

From these lexica we chose the NRC EmoLex lexicon, which originally has 14182 annotated utterances (words mostly). This was reduced to 1833 utterances according to an ad-hoc filtration method of our own device.

A row value of the data set always shows the utterance in the first column, and then a 0 or 1 annotation for anger, anticipation, disgust, fear, joy, sadness, surprise, and trust respectively. First the row values for each word were summed to see which utterances were identified by one single emotion annotation. From these, we selected the words that express only anger, fear, joy or sadness as these are the emotions present in System 1 (together with neutral). We already know from our NLP classification confusion matrices that some of these.

Then we presented the utterances that were annotated with 2 emotions and then cut the rows of the emotions we do not want to retain. From our previous classifications we know that e.g. disgust and fear and disgust and anger are difficult to keep apart. It may therefore even be a wise choice not to try to detect disgust at all. We don’t have the same experimental evidence for the other emotions that we cut, but we have no experimental information for these emotions from our System 1 experience and no way to evaluate.

Next, we selected all the utterances that were now indicated by only one annotation (having eliminated the second annotation through manual removal) and we added them to our list. Next we did the same for the utterances that were annotated with 3-4 emotions. It should be noted that in this last round only very few utterances passed the filter, however a significant number (252) of utterances in the category “joy” still came out of this round. Visual evaluation of the terms surprisingly revealed that most of the words were (to our subjective evaluation) very explicit in their expression of joy (rejoice, pleasant,…).

After this relatively unscientific filtering (however not devoid of logic) our lexicon now contains 1833 utterances each uniquely annotated for anger, fear, joy and sadness.

In order to do a check for possibly biased data, we downloaded the “bad\_words.txt” file from <https://www.cs.cmu.edu/~biglou/resources/> . This file contains words that could be considered offensive or profane by certain people. We checked our lexicon for these possibly offensive words, but we only took out the words that were annotated in what we felt was biased. The lexicon e.g. contained both the words “church” and “mosque” whereby church was annotated with happy and mosque was annotated with fear. We took out both words in this case.

Further words we took out on this basis were negro, lesbian, dike, black, wop, protestant, fat, fatty, sultan, aga and immigrant, which were all annotated with negative emotions. We removed them to avoid bias against certain demographics.

The utterances are modified by what we call “negation utterances” (because “I am not happy” contains the keyword “happy” but actually means the opposite). When a negation utterance is encountered the next keyword will be ignored.

Negation list: No, not, not a, not an, ain't, ain't a, ain't an, aint, aint a, aint an, wont, won't, will not, shant, shan't, shall not.

The system that we created is an extremely simple system with all its drawbacks and flaws. A state-of-the-art-system would have a deep analysis of the sentences and discourse for optimal word embeddings. This is what our BERT model does so well. It would also a create a knowledge representation that is domain and task specific. Our simple system does however have one big advantage: it is simple, and therefore fast.

Because the backchanneling has to be near-to-real time to make sense at all, and because the cloud-based STT system already takes some time to start producing text, the analysis of the text and the triggering of the backchanneling actions performed by the fish must be fast. The fish backchanneling is not supposed to be very “meaningful” it should show a reaction to the emotions of the person speaking, but it is up to the person to interpret the “mind” of the fish. Accuracy is therefore secondary to speed in this case.

# Potential Regulatory Issues for Privacy and Transparency

This part of the document was not set by the assignment, but because it was already clear from the beginning of the project that an AI application that detects and analyses people’s inner feelings will have serious privacy implications, this part was added for completeness’s sake. It is the result of a legal analysis regarding Emotion Detection in Smart Home Assistants. In view of the fact that our Goldfish of Solitude application fits that description quite well, the analysis is also applicable to the system we have created technically.

Apart from the identified privacy issues, there is proposed legislation from the European Commission to regulate certain AI use cases. The analysis in light of this proposed European “AI Act” brings forth even more serious considerations. That is why this analysis is included in this otherwise technical report.

* 1. *The applicable European regulatory framework.*

In 2018 *the EU Strategy on Artificial Intelligence*was published, which has three pillars, of which the *third pillar*is of relevance here, because the aim of this pillar is to *ensure an adequate ethical and legal framework for AI*. Within the context of the EU Strategy, the independent High Level Expert Group on Artificial Intelligence (the IS HLEG) was appointed.

2019 work by the AI High Level Expert Group was central to the development of the European approach to AI. The *7 key requirements*that were introduced by the *Ethics Guidelines*in the Assessment List for Trustworthy AI – *ALTAI* have guided the legislative steps. The 7 requirements were defined as:

1) Human agency and oversight; 2) Technical robustness and safety; 3) privacy and data governance; 4) Transparency; 5) diversity, non-discrimination, and fairness; 6) environmental and societal well-being; 7) accountability.

The AI HLEG’s work informed the *Communication on Building Trust in Human Centric Artificial Intelligence*, the *White Paper on Artificial Intelligence***,** and the *Coordinated plan on AI***.**

Then there is the work of the *OECD*, who with their *Recommendation* in 2019 not only created an important *definition of AI***,** butalso formulated *values-based AI Principles*around core issues. (Inclusive growth, sustainable development, and well-being – *Human-centered values and fairness – Transparency and Explainability*– Robustness, security and safety - Accountability).

In 2020 the *Council of Europe’s Ad-hoc Committee on Artificial Intelligence (CAHAI)*conducted a feasibility study and examined the reasons for why it is necessary to have an adequate legal framework for the protection of *human rights, democracy* ,and *rule of law*in light of the challenges posed by AI systems. This work is now followed (ongoing) by a multi-stakeholder consultation on these issues.

The European work has now culminated in a *legislative proposal* from the European Commission: *Proposal for a Regulation Laying Down Harmonized Rules on Artificial Intelligence*(Artificial Intelligence Act) – COM (2021) 206 final, issued on 21 April 2021. The document is a first of its kind, as it is the *first attempt at an AI law worldwide*, and its potential influence on world-wide AI policy making is therefore not negligible. It is however a *proposal* still and therefore nothing is yet cast in stone or executable. While the big lines of the framework will probably not change much, the details (such as e.g. definitions) will be fought over in the EP and Council (co-legislative procedure with compromise at the end). Notwithstanding its preliminary nature, this is the document to track to understand the thinking of the European legislator with respect to all things AI. More so, as the proposal is for a *Regulation*, which is a legal act of the EU that upon enactment *becomes immediately enforceable as law in all the Member States*, there is no requirement for transposition into national law.

We are however not there yet. The document is subject to proper legislative process, and this means that it will take at least two years before it (or better – an amended version) may become law. The first critiques are already rolling in, such as from BEUC, the European consumer group has already said that the Proposal “misses the mark” when it comes to consumer protection. Now compare this to the US FTC, which is all about consumer protection and treats all misrepresentation of AI technology, whereas the European proposal intends to regulate according to a *risk-based approach*, whereby certain minimal or no risk applications would be permitted without restrictions.

The proposed approach is *not construed as a regulation of AI technology, but regulates AI use cases* and is stepped as follows:

* **Unacceptable risk:** Prohibited
* **High risk:** Permitted subject to compliance with AI requirements and ex-ante conformity assessment by regulated 3rd party CE marking accreditation.
* AI with specific **transparency obligations:** Permitted but subject to information and transparency obligations.
* **Minimal or no risk:** Permitted without restrictions, but this is an exclusion scenario, which is only valid when all other risk levels have been assessed and excluded.

The proposal for an AI Act is the main legal framework we will be referring to when discussing the *Transparency aspects of our system*.

*10.2. Emotion Recognition in Smart Home Assistants in the light of Privacy/Data Protection and Transparency*

Both *Privacy and Data Protection*fall squarely under the moniker of Fundamental Rights as defined by the *Charter of Fundamental Rights of the European Union* (*2012/C 326/02*). These are expressed in Article 7 (respect for private and family life) and Article 8 (protection of personal data).

Article 16.1. of the *Treaty on the Functioning of the European Union* stipulates that: *Everyone has the right to the* protection of personal data*concerning them.*

Article 5.3. of the e-Privacy Directive says that Member States shall ensure that the use of electronic communications networks to store information or *to gain access to information stored in the terminal equipment of a subscriber or user is only allowed on condition that the subscriber or user concerned is provided with clear and comprehensive information in accordance with Directive 95/46/EC, inter alia about the purposes of the processing, and is offered the right to refuse such processing by the data controller*. This *shall not prevent any*technical storage or *access* for the sole purpose of carrying out or facilitating the transmission of a communication over an electronic communications network, *or as strictly necessary in order to provide an information society service explicitly requested by the subscriber or user*.

The e-Privacy Directive was meant for providers of electronic communications services (ECS). HAs are not ECS, they are information society services and that is why most of the e-Privacy directive is not applicable to HAs. apart from Article 5.3. It is unclear to which extent emotion detection algorithms would make use of information stored in the terminal device. This may not necessarily be applicable for all SHA implementations but when, for example, IP addresses or access to WiFi networks would be stored, the e-privacy Directive would become applicable in this sense. These data could be used to provide contextual data in an emotion detection system to ascertain the location of the HA. Expression of emotion is culturally dependent and such information could improve the performance of an ER system. It is therefore not implausible that such information would be retrieved and used.

The entire General Data Protection Regulation - GDPR is applicable in the case *personal data are collected and processed by automatic means*.

Whereby *‘****personal data****’ means any information relating to an identified or identifiable natural person (‘data subject’); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person;* (Article 4(1) GDPR)

This means that we cannot beforehand say whether HAs that perform emotion recognition are compliant. We have to analyze each and every use case independently**.** *It is not the technology that it is regulated, but its use.*

*10.3. Specific issues for Home Assistants in the light of the regulatory background*

Emotion Detection for Home Assistants form a *potentially grave intrusion into private and family life.*While users may consent to sharing the information that they knowingly transmit, they are mostly unaware that AI systems (including emotion detection systems) can *infer information* from the knowingly shared data. The SHA user may not actually want to share such inferred information.

Also, it is not really known which sounds activate the home assistant (apart from the default key words). The system can be activated unknowingly, and personal communication not intended for the SHA will also be shared. This potentially gives unwanted insights into family dynamics. This is not even touching upon the fact that to be keyword activated, these devices have to be *actively listening.* It is a small step from actively listening to surreptitious recording, which would make HAs effectively *spying machines*.

There is also the matter of the presence of children who may be *secondary users*of the device, invoking more issues, especially with respect to consent and profiling (as according to Article 8 GDPR and recital 38 of the GDPR).

Article 5(e) of the GDPR stipulates that personal data should not be stored after their purpose for which they have been collected has been served. Current Emotion Detection systems however are stooled on Machine Learning (ML) and ML requires Big Data. If the purpose of data collection is to create better emotion detection systems, the data are worth a fortune and will not be deleted until their purpose has been served (i.e. the creation of a “better” Emotion Detection system – which is an open-ended goal). This is why the “purpose” of collection has to be clearly and concisely defined.

In the case of Home Assistants *Biometric Data*are defined by Article 4(14) of the GDPR: *‘biometric data’ means personal data resulting from specific technical processing relating to the* physical***,*** physiological or behavioural characteristics of a natural person*, which allow or confirm the unique identification of that natural person, such* as facial images*or dactyloscopic data.*

All or at least one of these types of data are required to perform emotion detection. Emotion detection only from text (as obtained through STT transformation) is notoriously unreliable, as is emotion detection from facial imaging. Prosodic information (from the voice signal is most reliable of the three, but recently multimodal methods have shown most promise for more reliable emotion detection. This requires input from both the visual (facial input) as well as the speech channel with all the voice characteristics included (pitch, amplitude, etc.). *All these data can uniquely identify a natural person*. Processing of these data is in principle prohibited (Article 9.1 GDPR) unless the data subject gives explicit consent (Article 9.2(a)). It needs to be noted here that most consumers do not understand the extent of the technology and cannot give their consent in a sufficiently informed manner.

All this does not even take into account the fact that *secondary users* (including children) or *incidental users* (e.g. visitors) are not in a position to give or withdraw consent as they are not the “official” data subject.

Apart from these biometric data, other types of (non-personal) information can be collected that can help to infer emotional states (e.g. the type of music selected, type of movie selected, etc.) thereby conveying, what should in principle be internal emotional states.

Under *Article 22 of GDPR*, there is a specific prohibition to protect individuals against carrying out automated decision making that has legal or similarly significant effects on them. *This does not apply if the decision* is (amongst others) *based on*the data subject’s *explicit consent.*

This is a particularly tricky issue, as it is well known that people do not read Terms of Use and often will not even read popups before clicking “I consent”.

The *European Data Protection Board*adopted specific *Guidelines* on 9 March 2021 on *Virtual Voice Assistants.*VVAs are part of most if not all SHAs’ architecture. With respect to the issue that people don’t usually know what they are consenting to because they don’t read the messages, the Guidelines propose that VVAs should communicate in the form of auditory text, in order to make it easier for the consumer. But how would such a consent procedure work? Will people listen to what is said? Will they understand the full meaning of the request? Is a verbal acknowledgement binding? Should it be recorded? All difficult questions.

The VVA Guidelines further specify with respect to *privacy* that: *By* default, services which do not require an identified user should not associate any of the VVA identified users to the commands*. A privacy and data protection friendly default VVA would only process users’ data for executing users’ requests and would store neither voice data nor a register of executed commands.* Data controllers should, in general, apply *privacy by default and by design*principles. These instructions effectively render emotion detection impossible.

Lastly there is the question of *where the data will go* after they have been collected and processed? Will they perhaps be sold to e.g. medical insurance companies who want to assess the mental state of an individual in order to assess whether they will award life insurance? Is this at all desirable and do the potential benefits of emotion detection weigh up to concerning issues we have listed?

*10.4. Pros and cons of Emotion Detection in SHAs – The red lines of Privacy and Data Retention*

Let us assess the potential benefits. The main pros for emotion detection (in all AI driven systems, not just HAs) can be found in literature in the field of *“affective computing”*.

The main goal of *affective computing*is to develop systems capable of adapting users’ emotions to produce more natural and efficient interaction with machines. One of the motivations for the research is the ability to give machines *emotional intelligence*, including to *simulate* empathy.

Affective computing has applications in education, healthcare, video games, social monitoring and something called “affective mirrors”.

* *Education*

Affection influences learners' learning state. Computers could judge the learners' affection and learning state by recognizing their facial expressions. In distance education, due to the separation of time and space, there is no emotional incentive between teachers and students for two-way communication. Without the atmosphere brought by traditional classroom learning, students are easily bored, and this impacts the learning effect. Applying affective computing in distance education system is said to improve this situation.

* *Healthcare*

Social robots, as well as a growing number of robots used in health care benefit from emotional awareness because they can better judge users' and patient's emotional states and alter their actions/programming appropriately. This is especially important in those countries with growing aging populations and/or a lack of younger workers to address their needs.

Affective video games have been used in medical research to support the emotional development of autistic children.

* *Other applications*

A car could monitor the emotion of the people in the car and engage in additional safety measures, such as alerting other vehicles if it detects the driver to be angry.

Research and evangelization of affective computing technology can be found in the work of e.g. Daniel McDuff (MS) and Prof Rosalind W. Picard (MIT).

But when we search for “how can emotion detection/recognition improve privacy” we find nothing at all! As a matter of fact, nearly everything we find extolls the problematic character of this application of AI from a privacy point of view, including the points we ourselves have listed already.

The technology will more probably be used for personal profiling and targeted advertising than out of a desire to improve peoples’ lives. Taken one step further, it is the perfect way to assess how people can best be influenced, and this is not necessarily limited to their shopping behaviour, but can be extended to voting behaviour (as in the Cambridge Analytica incident) and any other behaviour as such, depending on the susceptibility of the individual.

Emotion detection is not an innocuous application of AI and everything depends on the use case and the exact goals and underlying implementation of the use case. We will go deeper into this in our section on Transparency, when we discuss the proposal for an AI Act.

Now we will look into the specifics of the Emotion Recognition technology as it exists today and evaluate an issue that we deem crucial, but that is not often spoken of: *can emotions, as inferred by AI be accurate?*

Article 5.1(d) GDPR stipulates that *Personal data shall be:* ***accurate*** *and, where necessary, kept up to date; every reasonable step must be taken to ensure that personal data that are inaccurate, having regard to the purposes for which they are processed, are erased or rectified without delay (‘accuracy’)*

As already explained previously*, all data that can uniquely identify a natural person* are considered personal data. ER uses biometric data for the inference process and the emotions are directly inferred from these biometric data. When the emotions are connected to personal data that identify a natural person, *they too become personal data*, which, according to Article 5.1 of the GDPR must be **accurate**.

Now let us take a look at how ER is done in practice:

Emotion detection is based on Emotion Models that define how emotions are represented. For emotion detection usually *discrete and/or discrete dimensional models*are used. Choosing to represent emotions discretely or dimensionally and not on a continuum has unavoidable consequences for the correctness of the analysis. *Representations always lump together various emotional states*as can be seen on Plutchick’s Wheel of Emotions from our introduction. This introduces a first *reductive bias* at the level of how we choose to *represent the reality* of a person’s true feelings and emotions.

*ER* on a large scale is usually done with *ML models* and these particular models mostly work on the basis of *supervised learning*, i.e. a technique that works with *labeled data*. Any Machine Learning process applied to labeled data sets *will only be as good as the labeling allows*. The way in which the feature set is established *has an impact on how data sets are labeled*. When human annotators are asked to label data (text, audio or video input) and they are offered only a *limited set of emotion labels*, there will be *reductive bias*in the labeling. If a Mechanical Turk respondent sees a video and believes the person in the video is jealous, but the only emotions that can be ticked for labeling are: anger, sadness, joy, fear, and neutral; the video will never be labeled for jealousy and the system will not learn this emotion, but will attach whatever label has been picked from the list of available emotions. This is a second bias, this time at the level of the data set.

Confusion matrices of ER systems show that *emotion pairs*such as anger/sadness, joy/neutral, shame/guilt, anger/disgust *can be easily confused*. Joy is the emotion detected best by vision and NLP systems. Incidentally, it is also the emotion that humans can recognize most accurately (which again suggests an -involuntary- *bias in the labeling of the data sets*). The label that is provided by a ML system for a particular emotion is not necessarily the correct one. This appears trivial, but what are the consequences for a person who is e.g. in an obligatory anger management program and who is being monitored, if anger and sadness are so readily confused?

*Multimodality*: humans are not very good at concluding emotions from facial features alone. We have at least facial, auditory, and textual (semantical, syntactical) inputs when we evaluate the emotional status of another person. Usually, we also have contextual clues and if we know the person, we also have information about their character and on how they usually behave. *Humans process information multimodally, ML models process information unimodally: NLP, computer vision and signal processing*being the main techniques that are used for ER and these three modalities can be inputted by an SHA system from samples taken by microphone and camera. *All three models separately have significant points of failure. Text* does not convey emotion very well: “yes” can be said in an angry, scared, happy,… tone. *Video* sampling rates (15 fps or higher) are insufficient to detect the minute and fleeting changes in expression that reveal human emotions, current demo’s always use almost clownish and frozen facial expressions for the good reason that the systems will not work on subtle (normal) expressions*. Signal processing* models so far are giving the “best” results, but these are mainly based on pitch (representing prosody) and amplitude (loudness), both quite limited representations of true emotion. There are research efforts ongoing to concatenate models in a multimodal way, but these models are experimental and far from ready for production.

The best NLP-based *Large language models such as BERT and GPT3 perpetuate biases contained in the training set data,* which are publicly available writing that is scraped from places such as Reddit and Wikipedia. If these texts contain biases, these biases will be captured and amplified in the output. Moreover, *the training sets are so large that the issues of bias in code cannot be properly documented, nor can they be properly curated to remove bias*.

It is also very well possible that by using large pre-trained models that are finetuned for specific tasks (such as ED), we are introducing a *pre-learning bias*and the system will learn less from our fine tuning than we might think. The consequence hereof is that the model *may not actually evaluate the text it is presented with for emotions,* but that it is presenting what it knows to be *“average” emotions learned in the pre-training phase.*

All this forces us to conclude that at this moment of technological development, it is virtually impossible to get emotion detection right and therefore the goal of accuracy is unattainable.

In fact, AI emotion detection is often referred to as a “pseudo-science” in the applicable literature.

Emotions are our deepest and most human essence. Emotions shape who we are and how we think of ourselves and others. And we may want to keep those opinions to ourselves, as is our right according to Article 11 of the European Charter of Human Rights. Are emotions really something we want to have assessed (constantly) by algorithms, especially when they are manifestly bad at it, without the option to be selective in when and what we share and what not?

*10.5. Applicable legislation with respect to transparency*

Since home assistants process personal data (e.g., users’ voice, location or the content of the communication), they must comply with the transparency requirements of the General Data Protection Regulation (GDPR) as regulated in *Article 5(1)(a) as well as Article 12 and Article 13 (enlightened by Recital 58).* Data controllers are obliged to inform users of the processing of their personal data in a *concise, transparent, intelligible form, and in an easily accessible way.*

Transparency is not defined in the GDPR. Nevertheless, *Recital 39 of the GDPR*informs about the meaning and effect of the principle of transparency in the context of data processing:

*“It should be* ***transparent*** *to natural persons that personal data concerning them are collected, used, consulted or otherwise processed and to what extent the personal data are or will be processed. The principle of* ***transparency*** *requires that* ***any information and communication relating to the processing of those personal data be easily accessible and easy to understand****, and* ***that clear and plain language be used****. That principle concerns, in particular, information to the data subjects on the* ***identity of the controller*** *and* ***the purposes of the processing*** *and further information to ensure fair and transparent processing in respect of the natural persons concerned and their* ***right to obtain confirmation and communication of personal data concerning them which are being processed****…”*

The *Article 29 Working Party Guidelines* provide *practical guidance and interpretative assistance on the obligation of transparency concerning the processing of personal data under the GDPR.*

The *Assessment List for Trustworthy Artificial Intelligence (ALTAI)* includes *recommendations* from experts in the domain of AI (the High-Level Expert Group on AI). The ALTAI states that *transparency* is a crucial component for achieving *Trustworthy AI*and divides the matter into three subcategories: *traceability, explainability, and open communication about the limitations of the AI system.*

The European Data Protection Board (EDPB) *Guidelines on Virtual Voice Assistants* (VVAs) identify some of the most relevant challenges for VVAs (such as complying with *transparency*, privacy, security, and data regulations included in the GDPR and other legislations) and provide *recommendations* to relevant stakeholders on how to address them.

*In view of the fact that most SHA incorporate VVA technology, these guidelines are also relevant for our analysis.*

Given the specific nature of SHAs, data controllers face several obstacles to comply with the *GDPR’s* transparency requirements:

* *multiple users:* all users (registered, non-registered, and accidental users) should be informed about the use of their data, not only the user setting up the home assistant.
* *ecosystem complexity:* the identities and roles of those processing personal data when using a home assistant is far from evident for the users.

The EDPB highly recommends that users should be informed by the SHA device of the status in which it currently lies – e.g., whether the SHA is currently listening to them or recording what they are saying. This enhancement can be achieved by designers and developers by *making the SHA more interactive* (e.g., the device might acknowledge in some way the reception of a vocal command) and *broadcasting the status* of the machine with specific signals. There are many options that can be explored in this regard, ranging from the use of specific vocal acknowledgements and visible icons or lights, or the use of displays on the device.

Two important questions arise from the aforementioned recommendation: 1) what is it the *most feasible way to inform users* and 2) *when is the appropriate time to inform them?* It is also necessary to consider whether the SHA is used by *one user* or potentially *multiple users as is more habitual for a SHA.*

Sometimes SHA designers require the user to create accounts which bundle the assistant’s service with multiple other services (like email, video streaming, and purchases) and applications. The decision by the SHA designer to link the account to other services may have the effect of requiring very lengthy and complex privacy policies. *Privacy policies of great length and complexity should be avoided as they hinder the fulfilling of the transparency principle.*

Currently, all SHAs, are connected to a user account or set up by an application that requires one. The question of how data controllers inform these users about the privacy policy while setting up the SHA should be addressed as described in the *Article 29 Working Party Guidelines*. The SHA providers and all applications providers should make the necessary *information available prior to registration or downloading of new apps*. This way the information is given at the earliest possible time, and at the time when the personal data is obtained. SHA’s using this third-party app deployment strategy should ensure that users get the necessary information also on the third-party processing.

In order to comply with the GDPR, data controllers should find a way *to inform not only registered users, but also non-registered users and accidental SHA users.* This condition could be especially difficult to fulfil in practice, especially in the case of *children, elderly people*, and *people with impairments or disabilities*. Adequate information should make the data subjects understand whether their use of the *SHA will be linked to other processing activities* managed by the service provider apart from the strict use of the SHA, *including the fact whether there are any emotion detection algorithms at work.*

We have not found any concrete examples that would make ER or even SHAs more transparent, but it is evident that governments are already driving towards the positive use of AI systems. As an example, the Italian data protection authority has recently introduced new requirements on the video-sharing app TikTok in a bid to protect underage users. The new requirements for TikTok are to guarantee the deletion of underage accounts within 48 hours, to draw up a new information notice for young users, and to develop *AI solutions to detect underage users*. It is clear that industry will have to work towards improving the transparency also in the area of SHAs and the use of emotion recognition.

One example that springs to mind but which is not directly connected to emotion detection is Apple’s recent initiative to allow the users of their devices to limit the personalization of ads and how to turn off location-based ads. While such option would increase transparency (and even control) for end-users, it is quite clear that there are questions pertaining to competition law on whether this does or does not constitute an abuse of market power. In the weighing between transparency and competition, it is not unlikely that the balance could be struck in favour of competition (to the detriment of transparency and control for the end-user).

Overall, the use of ER introduces potential risks and issues in terms of the transparency of SHAs.

*Emotion recognition is a specific task:*

The task of ER in practice does not use statistical machine learning modes (like Random Forest) which are to some degree human-readable and understandable. Emotion detection is usually performed by means of Deep Neural Networks (DNNs) and these architectures are becoming more and more complex. AI technology surpasses our ability to reason and understand often making it impossible for humans to know why an AI system makes certain decisions even with the most advanced techniques of record-keeping and analysis. The use of ER is likely to further push untransparent systems into application. There are (research) efforts underway to make such systems more transparent, but we are currently at a point where the performance of DNNs is vastly better than any other algorithmic approach, whereas the techniques to make these algorithms more transparent are only nascent.

*Informing end-users:*There are several issues in this regard. A SHA is bought by a single person but might be used by multiple members of the household. It is therefore difficult to ensure that they have all been informed about the use of the data. The traditional way of informing end-users by including the necessary information in a manual is not sufficient to guarantee that everyone using the device is aware of how their data is being processed, stored, and used. Manuals are rarely read and if they are read, then usually only by the person installing the system.

*Providing end-users access to their data:*In case a user requests access to the information that an emotion recognizer has stored (as these are personal data) there is the question of how to provide such information in a transparent matter. ER systems constantly monitor and keep track of inputs and storing the emotions that go with every utterance implies storing massive amounts of data. It is an open question how and in which format such data should be provided to the requesting end-user and if such massive amount of data can actually be transparently read, interpreted and understood by the end-user. Additionally, in the case where accidental users who have not registered with the SHA device interact with it and have their emotions recognized, it may be impossible or extremely difficult for them to request access to their data. This does not even address the matter whether accidental users will be recognized by the system as such or if their data will be stored under the account and the identity of the owner of the SHA.

*10.6. Transparency in light of the Proposed AI Regulation:*

The European Commission published a Proposal for an AI Regulation on 21 April 2021. The proposed regulation introduces a comprehensive regulatory framework for AI. The ambition is to provide the legal certainty needed to facilitate investment and innovation in AI, whilst establishing a framework to safeguard fundamental rights and ensure safe use of AI applications. The Proposal introduces specific regulation for all AI systems, including for technologies such as ER and SHAs.

The new legislation will follow a risk-based approach in deciding how to regulate different AI systems:

* **Prohibited:** AI systems which are considered a *clear threat to the safety, livelihoods and rights of people will be banned*. Such systems include AI systems or applications that manipulate human behaviour to circumvent users' free will.   
  Article 5 stipulates (amongst others) that: *the placing on the market, putting into service or use of an AI system that* deploys subliminal techniques beyond a person’s consciousness in order to materially distort a person’s behaviour in a manner that causes or is likely to cause that person or another person physical or psychological harm*.*
* **High-risk:** AI systems that *pose significant risks to the health and safety or fundamental rights of persons.* Article 6.2 refers to an Annex III where some of such high-risk systems are listed and Point 1. of this Annex III specifies that: *“Biometric identification and categorisation of natural persons: (a) AI systems intended to be used for the ‘real-time’ and ‘post’ remote biometric identification of natural persons.”*
* **Low risk:** AI systems that are considered sufficiently low only get specific transparency obligations according to Article 52: “Providers shall ensure *that AI systems intended to interact with natural persons* are designed and developed in such a way that *natural persons are informed that they are interacting with an AI system, unless this is obvious from the circumstances and the context of use.*”
* **No risk:** Systems that are not considered prohibited, high risk or low risk the fall into the category of no-risk, but this is not a defined category. Every system or application will have to be evaluated on a case-by-case basis and if they pass the sieve of the prohibited, high-risk and low-risk categories, only then can they be considered no risk.

*For all the categories there are Transparency obligations to abide by.* The Explanatory Memorandum to the Proposal clarifies that transparency obligation in general apply for systems that **interact with humans**, that are used to **detect emotions**, or are based on **biometric data**. When people interact with an AI system or their emotions or characteristics are recognised through automated means, people *must be informed* of that circumstance. All of which applies to the SHA systems we are discussing here today. The Proposal is currently just that: a first proposal. Not everything is clear or clearly defined yet, the document will crystallise in the course of the legislative process. It is therefore sometimes difficult to interpret at this point, but that should not keep us from thinking of ER on SHAs in the light of the existing text.

With respect to the Prohibited category of AI systems, we should ask ourselves whether there are any *use cases* of ER on SHAs that could consist of “*subliminal techniques beyond a person’s consciousness”* of which the “*aim*” is to “*materially distort a person’s behaviour*” in a “*manner that causes or is likely to cause physical or psychological harm*”. While at first glance this seems to be a tall order, we will look at the seemingly innocuous example of a *music recommender system such as Spotify, to which users can have access through SHAs.* The aim of such a system is to persuade its users to buy more music, as this is their business case.

Could emotion detection be considered a “subliminal technique beyond a person’s consciousness”. We think it could, especially if the end-user is unaware that emotion detection is used. The case becomes murky when the *person is informed* of the fact that the recommender system uses emotion detection. If this fact is communicated once, to the person installing the system, is this *sufficiently transparent* for all users of the SHA to become conscious at all times that the device is using subliminal techniques to push its music?

The question whether the system’s *aim* is to *materially distort a person’s behaviour* is even more difficult to answer. We are looking at intentionality here. How can we know what the intention of using such system is? Playing Devil’s Advocate, we could say that the recommender system wants to please the end-user and offer a better service but knowing the business case is to sell more music, we can actually assume that the aim of such recommender system is to sell more music and sometimes even to sell certain kinds of music e.g., where the music platform and the artist have a deal that is in favour of the platform. So, in fact we could say that the *aim is to distort the end-user’s behaviour* and to make them buy more and specific music.

Lastly, could the use of the technique cause *physical or psychological harm* (forbidden as per Article 3 CFEU). At the risk of sounding overly dramatic, we can imagine instances where this could actually be the case. Music recommender systems are there to sell music. They are not psychologists or psychiatrists, and they have no idea what *the consequences could be of reinforcing certain emotions* through the selective pushing of certain types of music. If a person signals that their emotional stat is “sad” and their music preference at the time is in the “depressive” register, a recommender system could very well conclude that the person is in a sad mood and wants to listen to sad music, whereas the person may be fundamentally and clinically depressed and reinforcement of such depression, especially by music, which is known to have a big influence on mood, *may have very dire**consequences* not only on the *psychological* condition of the person but in extreme case also on their *physical* integrity.

Physical or psychological damages are difficult to prove as per tort law, the recommender system owners could always claim that their intentions were pure. But when we regard such claim from the perspective of *Bentham’s Consequentialist ethics* principles, this cannot be an excuse: *even if we didn’t know there would be bad consequences and we had the best intentions in the world, we are still responsible for those consequences.*

In brief: there are use cases that can be considered red line cases, and it is *unclear how we can mitigate* the red line through *transparency measures*. How can we make sure that the end-user of the system is conscious *at all times* of the techniques used on them?

The Proposal for an AI Act has taken the principles of the HILEG’s Ethics Guidelines for Trustworthy AI (ALTAI) into account and these principles are woven into the fabric of the proposal, also with respect to the principle of transparency.

We will now look at the Proposal with respect to transparency in terms of the three compounding elements as defined in the ALTAI: explainability, traceability, and communication about the AI system’s limitation and capabilities.

*Explainability***:**

According to the ALTAI explainability concerns the ability to *explain both the technical processes of an AI system and the related human decisions.* Technical explainability requires that the decisions made by an AI system *can be understood and traced by human beings*. Moreover, trade-offs might have to be made between enhancing a system's explainability (which may reduce its accuracy) or increasing its accuracy (at the cost of explainability). Whenever an AI system has a significant impact on people’s lives, it should be possible to demand a *suitable explanation of the AI system’s decision-making process*. Such explanation should be *timely* and *adapted to the expertise of the stakeholder concerned*. In addition, explanations of the degree to which an AI system influences and shapes the organisational decision-making process, design choices of the system, and the rationale for deploying it, should be available (hence *ensuring business model transparency*).

When using ER in home assistants for the detection of the emotional/mental state of a person it is likely that the efficiency of the emotion recognizer will come at the cost of its explainability. This is because the most accurate ER models are based on deep learning models (of which most are blackbox) that are not conducive for humans to trace and understand the decisions of such systems.

Even when technical persons do have an understanding of the decision-making process such as is the case for probabilities and statistics-based systems, it is not evident to adapt the explanations of decisions made by AI systems to the level of the expertise of different.

According to *Annex 3.1(a)* of the Proposal, systems that *allow biometric identification and categorisation of natural persons and can be used for “real-time” and “post” remote biometric identification of natural persons are considered high-risk*. In view of the fact that ER in SHAs uses technology that can be interpreted to be “Biometric identification and categorisation of natural persons*”* our system, if not categorised as Prohibited, could in any case be categorised as high-risk according to the Proposal.

*High-risk AI systems* must be *developed* in a way to enable the *automatic recording of events* (“logs”) as covered in Article 12 of the Proposal. The logging capabilities must enable the *monitoring of the operation* of the high-risk AI system with respect to the *occurrence of situations that may result in the system presenting a risk and conform to recognised standards* or common *specifications*, providing at a minimum:

* recording of the period of each use of the system.
* reference database against which input data has been checked by the system.
* identification of the people involved in the verification of the results.

Furthermore, it is necessary for such high-risk systems to be designed and developed in a way to ensure their operation is sufficiently transparent to enable stakeholders to interpret the system’s output. Therefore, SHAs must be accompanied by instructions for use that include concise, complete, correct, and clear information that is relevant, accessible, and comprehensible to stakeholders (Article 13), such as:

* identity and the contact details of the provider.
* characteristics, capabilities and limitations of performance of the high-risk AI system, including:
  + its intended purpose.
  + the level of accuracy, robustness and cybersecurity.
  + any known or foreseeable possible misuse, which may lead to risks to the health and safety or fundamental rights.
  + its performance as regards to the persons on which the system is intended to be used.
  + when appropriate, specifications of the data used for training.
* any the changes to the system and its performance.
* human oversight measures put in place to facilitate the interpretation of the outputs of AI systems by the users.
* expected lifetime of the high-risk AI system and any necessary maintenance measures to ensure its proper functioning.

*Communication***:**

The ALTAI tells us that AI systems should *not represent themselves as humans to users*; humans have the *right to be informed that they are interacting with an AI system*. This entails that AI systems must be identifiable as such. In addition, the option to decide against this interaction in favour of human interaction should be provided where needed to ensure compliance with fundamental rights. Beyond this, the *AI system’s capabilities and limitations should be communicated to AI practitioners or end-users* in a manner appropriate to the use case at hand. This could encompass *communication of the AI system's level of accuracy, as well as its limitations*.

The considerations from the ALTAI that could be applicable to our emotion detecting SHA are treated in Article 52 (*Transparency obligations for certain AI systems)*, subsections 1 and 2 of the Proposal:  
*Article 52.1. Providers shall ensure that AI systems* intended to interact with natural persons *are designed and developed in such a way that natural persons are informed that they are interacting with an AI system, unless this is obvious from the circumstances and the context of use.*

Article 52.2. *Users of an emotion recognition system or a biometric categorisation system shall inform of the operation of the system the natural persons exposed thereto.*

The lawmakers had specific applications (such as chat bots) in mind when they wrote the lines in Article 52.1. They believe that certain AI systems intended to interact with natural persons may pose specific risks of impersonation or deception irrespective of whether they qualify as high-risk or not.

The text of the article is however quite broad and does not include the impersonation risks which are only mentioned in the recitals. It is therefore at this point *unclear whether a SHA would be considered as “interacting with natural persons” (we would say yes) and whether it is obvious from the circumstances and the context of use in the case of SHAs.*

With home assistants, it is hard to guarantee that everyone living in the household is aware at all times that they are interacting with an AI system.

How and when should end-users be informed? Should there be a visual and/or audial signal every time people interact with the SHA to indicate that they are interacting with an AI, or is a one-time warning sufficient? (We imagine a product packaging with a label “AI Inside” here). Do any of these practices suffice for the elderly, people with hearing or visual impairments or people with disabilities?

This question is particularly an issue with children as it is difficult to ensure they understand what AI is in the first place and the purposes for which e.g., the recognition of their emotions could be used. It is difficult to ensure that all members of the household are informed on an efficient basis of these purposes.

Lastly, it appears that the Article 52 obligations might be cumulative with other obligations, meaning that *if a use case is considered high-risk AND interacting with natural persons, the transparency obligations for both the high-risk systems and those imposed on AI systems interacting with natural persons would apply.*

Another issue is how to inform people – should there be a visual and/or audial signal every time people interact with a SHA to indicate that they are communicating with an AI? Does this suffice for the elderly, people with hearing or visual impairments or people with disabilities?

*Traceability***:**

Lastly, the ALTAI says with respect to traceability that the *data sets* and the *processes that yield the AI system’s decision*, including those of *data gathering and data labelling* as well as the *algorithms used*, should be *documented to the best possible standard* to allow for traceability and an increase in transparency. This also applies to the *decisions made by the AI system*. This enables *identification of the reasons why an AI-decision was erroneous* which, in turn, could help prevent future mistakes. *Traceability facilitates auditability as well as explainability*.

With respect to the traceability of the *data sets and processes* that yield the AI system’s decisions, there is an issue with *tracing back the decision-making of the system*, even if we diligently log all the data sets used for training and testing and all the processes (algorithms).

In terms of *data gathering and labelling* traceably implies that provenance of the data should be *known and qualitatively sound*. This is not always the case as data is often *scraped off the public internet*. Even publicly available and well documented data sets suffer from issues we have already discussed in the section on accuracy of emotion detection in that the *annotation is unavoidably biased*. The same is also the case for *algorithms* used to perform emotion detection and the *decisions* made by the AI system.

It is crucial for SHAs providers to keep records and have available technical documentation which contains information that is necessary to assess the compliance of the AI system with relevant requirements. This is even more so the case when the SHA uses emotion detection and thus, as mentioned earlier, presents an increased risk on the fundamental rights and safety of end-users.

For high-risk AI systems, the requirements of *high-quality data* and *documentation* are strictly necessary to mitigate the risks to fundamental rights and safety posed by AI and that are not covered by other existing legal frameworks. In terms of record-keeping there is also the issue of what meaningful data to log. Having information on how high-risk AI systems have been developed and how they perform throughout their lifecycle is essential to verify compliance (Article 11). Therefore, up-to-date information needs to be kept on the following:

* technical documentation necessary to assess the compliance of the AI.
* general characteristics, capabilities and limitations of the system, algorithms, data, training, testing, and validation processes.
* documentation on the relevant risk management system.

In the case of emotion detection, we conclude that even the most diligent tracing of all the steps involved *may not lead to the identification* of the reasons why an *AI-decision was erroneous* and why transparency measures may be insufficient to mitigate the risks for ER systems. This leads us to conclude that emotion detection systems whether or not implemented in SHAs may present a number of red lines begging the question whether this use-case is desirable in the light of transparency and privacy requirements.

*10.7. Mitigating factors for the Goldfish of Solitude project*

Clearly, as long as we do not deploy this system, it does not have to adhere to all these rules. Nevertheless, because we were aware even from the beginning that any emotion detection system could have serious privacy issues there were certain mitigating factors that we built in from the go.

The fact that we intended the recipient of the emotion detection reporting to be a person with whom the end-user of the system already has a regulated confidential relationship (such as a doctor or psychologist) and that the data are not to be sent elsewhere for further processing is already a major mitigating factor with respect to privacy concerns.

Also, the fact that we have diligently kept track of the data sets and methods used, as well as the potential biases that we can see creeping into the system comply with some of the transparency requirements set for high-risk AI systems.

Even with these mitigating factors, we believe that while this was an excellent subject for a project, deploying such a system in reality would be fraught with serious legal and regulatory issues, especially once the proposed AI Act will come into force.

Bibliography (legal section)

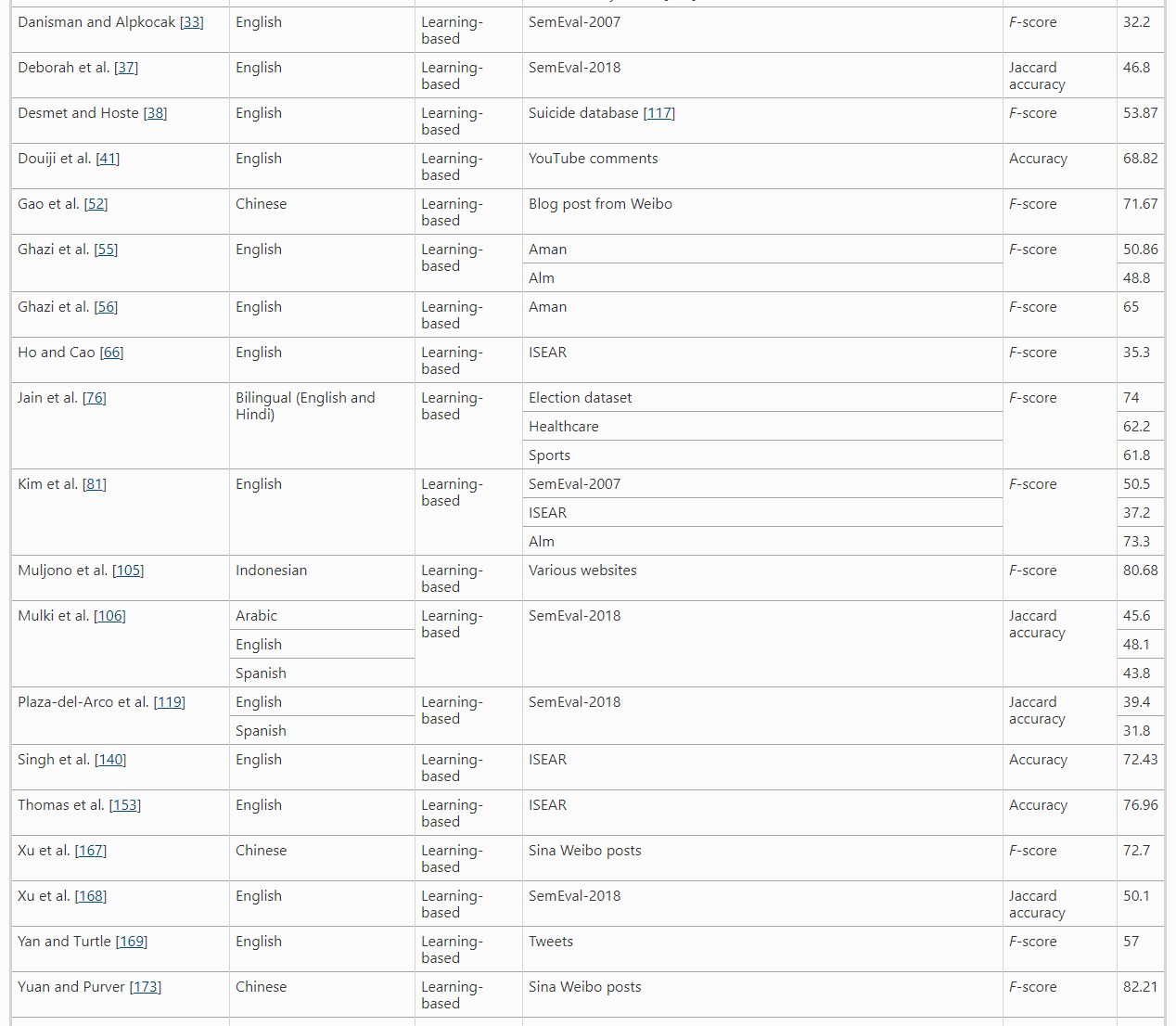
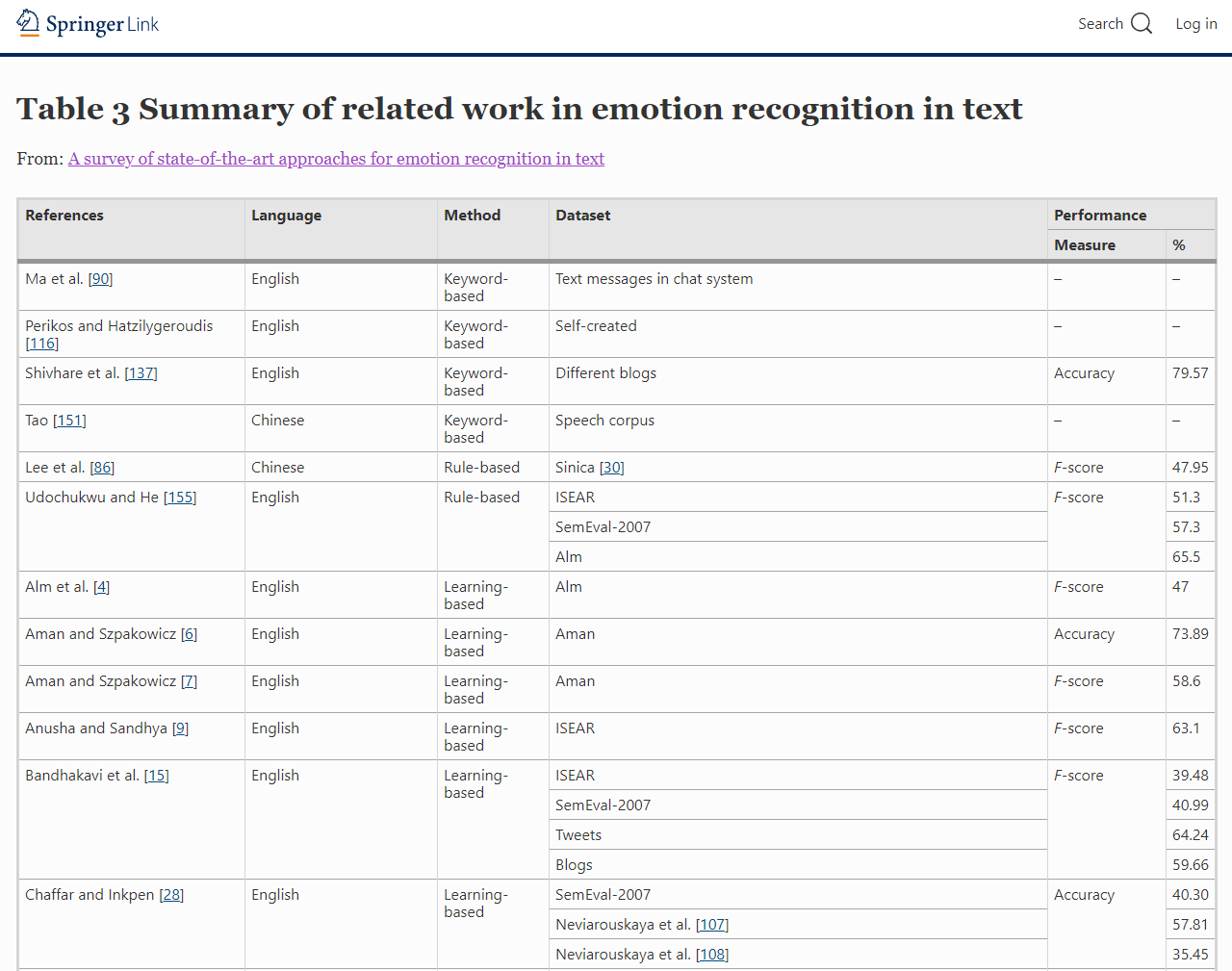
* World Economic Forum: <https://www.weforum.org/projects/global-ai-action-alliance>
* Federal Trade Comission: <https://www.ftc.gov/news-events/blogs/business-blog/2021/04/aiming-truth-fairness-equity-your-companys-use-ai>
* CAHAI
  + <https://www.coe.int/en/web/artificial-intelligence/cahai>
  + <https://rm.coe.int/prems-107320-gbr-2018-compli-cahai-couv-texte-a4-bat-web/1680a0c17a>
  + <https://forms.office.com/pages/responsepage.aspx?id=grBJPtViSUilsIbtUZKH0qerRpgKeM5LqbRG77gIXQhUMklQV1JEUEFRQ1hKUDZaOEc1WDc1TVNPSi4u>
* OECD
  + <https://www.oecd.ai/ai-principles>
  + <http://mneguidelines.oecd.org/RBC-and-artificial-intelligence.pdf>
  + <https://oecd.ai/assets/files/OECD-LEGAL-0449-en.pdf>
* AI –HLEG ALTAI: <https://ec.europa.eu/digital-single-market/en/news/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment>
* EU Artificial Intelligence Act (proposal): <https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-european-approach-artificial-intelligence>
* EU Charter Fundamental Rights: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:12012P/TXT>
* Treaty on the Functioning of the European Union (TFEU): <https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:12012E/TXT:en:PDF>
* General Date Protection Regulation (GDPR): <https://eur-lex.europa.eu/eli/reg/2016/679/oj>
* E-Privacy Directive: <https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32002L0058>
* Guidelines 02/2021 on Virtual Voice Assistants: <https://edpb.europa.eu/our-work-tools/documents/public-consultations/2021/guidelines-022021-virtual-voice-assistants_en>
* Guidelines on transparency under Regulation 2016/679 (Article 29 Working Party guidelines on transparency) : <https://ec.europa.eu/newsroom/just/document.cfm?doc_id=48850>
* Websites, blogs:
  + <https://www.kaspersky.com/blog/emotional-ai/31689/>
  + <https://www.louroe.com/product/aggression-detector/>
  + <https://www.mobihealthnews.com/content/study-finds-new-program-using-google-glass-ai-helps-children-autism-interpret-emotions>
  + <https://www.gartner.com/smarterwithgartner/13-surprising-uses-for-emotion-ai-technology/>
  + <https://news.engin.umich.edu/2020/02/emotion-recognition-has-a-privacy-problem-heres-how-to-fix-it/>
  + <https://www.zdnet.com/article/ethics-of-ai-the-benefits-and-risks-of-artificial-intelligence/>
  + <https://support.apple.com/en-us/HT202074>
  + <https://nypost.com/2021/01/23/italy-orders-tiktok-to-block-accounts-for-kids-after-girls-death/>
* Video:
  + Responsible Tech Podcast #3 - Kamervragen aan Grapperhaus en Artificial Intelligence Act: <https://www.youtube.com/watch?v=1DJlUDUr17k>
  + Emotion aware technology - improve well-being and beyond | Daniel McDuff | TEDxBerlin: <https://www.youtube.com/watch?v=9iQiGpSmQ74>
  + What If Your Emotions Were Tracked To Spy On You? (European Parliamentary Research Service Blog): <https://epthinktank.eu/2019/03/14/what-if-your-emotions-were-tracked-to-spy-on-you-science-and-technology-podcast/>
* Lectures:
  + vander Maelen, C. and Vermeulen, J: Data protection and AI: The case of smart home assistants (May 5, 2021): <https://www.law.kuleuven.be/citip/en/news/item/conference-artificial-intelligence-and-the-law-may-5-2021>
* Patent applications and Patents:
  + <https://worldwide.espacenet.com/patent/search?q=pn%3DUS2014280296A1>
    - Google: Providing help information based on emotion detection
  + <https://worldwide.espacenet.com/patent/search?q=pn%3DUS2017140049A1>
    - IBM: Web search based on browsing history and emotional state
* Articles, books:
  + Kim, D., Ahn, S., Park, S., & Whang, M. (2013). Interactive emotional lighting system using physiological signals. IEEE Transactions on Consumer Electronics, 59(4), 765–771
  + Rincon, J. A., Costa, A., Carrascosa, C., Novais, P., & Julian, V. (2019). EMERALD-Exercise Monitoring Emotional Assistant. Sensors, 19(8), 1953.
  + Park, J.-S., Kim, J.-H., & Oh, Y.-H. (2009). Feature vector classification-based speech emotion recognition for service robots. IEEE Transactions on Consumer Electronics, 55(3), 1590–1596.
  + Plutchik, R. (2001). The Nature of Emotions. American Scientist, 89(4), 344.
  + Lövheim, H. (2012). A new three-dimensional model for emotions and monoamine neurotransmitters. Medical Hypotheses, 78(2), 341–348.
  + Ekman, Paul. *Emotion in the Human Face*. 2nd ed., Cambridge University Press, 1982.
  + Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements: Psychological Science in the Public Interest, 20(1), 1–68.
  + Imbir K.K. (2017) Psychoevolutionary Theory of Emotion (Plutchik). In: Zeigler-Hill V., Shackelford T. (eds) Encyclopedia of Personality and Individual Differences. Springer, Cham. <https://doi.org/10.1007/978-3-319-28099-8_547-1>
  + Van Gool, E., De Bruyne, J. ; Gils, T., [Tort Law and Damage Caused by AI Systems](https://lirias.kuleuven.be/3393637), 2021-02, pp 359 - 403, ISBN: 9781839701030, Intersentia
  + Furey, E., & Blue, J. (2018). Alexa, Emotions, Privacy and GDPR, Conference paper, Human Computer Interaction Conference, July [2018].
  + McDuff, D. J., Rowan, K., Choudhury, P., Wolk, J., Pham, T., & Czerwinski, M. (2019). A Multimodal Emotion Sensing Platform for Building Emotion-Aware Applications. ArXiv Preprint ArXiv:1903.12133.
  + Borod JC. (2000). The Neuropsychology of Emotion. Oxford, UK: Oxford University Press.
  + Jaiswal, M., & Provost, E. M. (2020). Privacy Enhanced Multimodal Neural Representations for Emotion Recognition. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, pp. 7985–7993).

Annex 1: Benchmarks

**State of the Art[[23]](#footnote-23)**

**TABLE 2.**Summary of current advances in text‐based emotion detection[[24]](#footnote-24)

| **Proposed Work** | **Approach** | **Dataset(s)** | **tion** | **Limitations** |
| --- | --- | --- | --- | --- |
| Ahmad et al, 2020[90](https://onlinelibrary.wiley.com/doi/10.1002/eng2.12189?fbclid=IwAR2spFkZuRYFGN5sJOoDtlhz1AVbUskUU7QqEy4Jvi0doGZMMPEWPxdoYdU#eng212189-bib-0090) | Machine Learning | Emo‐Dis‐HI data | Showed that knowledge gathered from resource‐rich languages can be applied to other language domains using transfer learning and cross‐lingual embeddings. Obtained an F1 score of 0.53 | • Disregard for the contextual meaning of words. |
| Matla and Badugu, 2020[91](https://onlinelibrary.wiley.com/doi/10.1002/eng2.12189?fbclid=IwAR2spFkZuRYFGN5sJOoDtlhz1AVbUskUU7QqEy4Jvi0doGZMMPEWPxdoYdU#eng212189-bib-0091) | Machine Learning | Tweets | Showed the efficiency of the NB in comparison with the KNN machine learning technique, with an accuracy of 72.06*%* to 55.50*%*. | • Low extraction of contextual information in sentences. |
| Seal et al, 2020[65](https://onlinelibrary.wiley.com/doi/10.1002/eng2.12189?fbclid=IwAR2spFkZuRYFGN5sJOoDtlhz1AVbUskUU7QqEy4Jvi0doGZMMPEWPxdoYdU#eng212189-bib-0065) | Rule Based | ISEAR data | Detected emotions with special attention to phrasal verbs in the ISEAR dataset | • Disregard for the contextual meaning of words. |
|  |  |  |  | • Inadequate vocabulary of words in the lexicon. |
| Huang et al, 2019[89](https://onlinelibrary.wiley.com/doi/10.1002/eng2.12189?fbclid=IwAR2spFkZuRYFGN5sJOoDtlhz1AVbUskUU7QqEy4Jvi0doGZMMPEWPxdoYdU#eng212189-bib-0089) | Machine Learning | EmotionLines dataset | Investigated the performance of BERT on the EmotionLines dataset. Their model attained an F1 Score of 0.815 and 0.885 for the Friends and EmotionPush datasets respectively. | • Inadequate amount of data |
| Huang et al, 2019[87](https://onlinelibrary.wiley.com/doi/10.1002/eng2.12189?fbclid=IwAR2spFkZuRYFGN5sJOoDtlhz1AVbUskUU7QqEy4Jvi0doGZMMPEWPxdoYdU#eng212189-bib-0087) | Machine Learning | Tweets | Presented an HRLCE and BERT model for text‐based emotion detection. They attained an F1 Score of 0.779 for happy, angry and sad emotion classes. | • High number of misclassifications. |
| Ragheb et al, 2019[81](https://onlinelibrary.wiley.com/doi/10.1002/eng2.12189?fbclid=IwAR2spFkZuRYFGN5sJOoDtlhz1AVbUskUU7QqEy4Jvi0doGZMMPEWPxdoYdU#eng212189-bib-0081) | Machine Learning | SemEval‐2019 Task 3 dataset | Presented an attention based model for categorizing emotions. They attained an F1 Score of 0.7582 | • Does not perform well for detecting the happy emotion. |
| Malte and Ratadiya, 2019[86](https://onlinelibrary.wiley.com/doi/10.1002/eng2.12189?fbclid=IwAR2spFkZuRYFGN5sJOoDtlhz1AVbUskUU7QqEy4Jvi0doGZMMPEWPxdoYdU#eng212189-bib-0086) | Machine Learning | Facebook multilingual texts | Proposed a bi‐directional transformer BERT architecture. Attained an F1 score of 0.4521 for Hindi texts and 0.5520 for English texts | • The presence of slangs in the non‐English text hindered system's performance. |
| Polignano et al, 2019[80](https://onlinelibrary.wiley.com/doi/10.1002/eng2.12189?fbclid=IwAR2spFkZuRYFGN5sJOoDtlhz1AVbUskUU7QqEy4Jvi0doGZMMPEWPxdoYdU#eng212189-bib-0080) | Machine Learning | ISEAR dataset, the SemEval‐2018 Task1 dataset and the SemEval‐2019 Task 3 dataset | Presented a word embedding method for text based emotion detection by comparing the performance of the Google Word Embedding, GloVe Embedding and the FastText Embedding. FastText embedding performed better | • proposed model has a high complexity. |
| Huang et al, 2019[56](https://onlinelibrary.wiley.com/doi/10.1002/eng2.12189?fbclid=IwAR2spFkZuRYFGN5sJOoDtlhz1AVbUskUU7QqEy4Jvi0doGZMMPEWPxdoYdU#eng212189-bib-0056) | Machine Learning | SemEval | Showed that EMN + SA yields better results than CNN, TLSTM, and RecNN used individually | • Not good for generalization due to limited number of categories |
| Ma et al, 2019[79](https://onlinelibrary.wiley.com/doi/10.1002/eng2.12189?fbclid=IwAR2spFkZuRYFGN5sJOoDtlhz1AVbUskUU7QqEy4Jvi0doGZMMPEWPxdoYdU#eng212189-bib-0079) | Machine Learning | SemEval | Used the Bi‐LSTM to classify emotions in textual and emoji utterances into happy, sad and angry and detected that their system's performance for happy and angry outperformed baseline models but not for sad. They concluded that Bi‐LSTMs are capable of extracting contextual information from texts. | • Restricted categories of emotion classes |
| Chatterjee et al, 2019[58](https://onlinelibrary.wiley.com/doi/10.1002/eng2.12189?fbclid=IwAR2spFkZuRYFGN5sJOoDtlhz1AVbUskUU7QqEy4Jvi0doGZMMPEWPxdoYdU#eng212189-bib-0058) | Machine Learning | SemEval | Detected four classes of emotions using the LSTM | • Not good for generalization due to limited number of categories |
| Alotaibi, 2019[52](https://onlinelibrary.wiley.com/doi/10.1002/eng2.12189?fbclid=IwAR2spFkZuRYFGN5sJOoDtlhz1AVbUskUU7QqEy4Jvi0doGZMMPEWPxdoYdU#eng212189-bib-0052) | Machine Learning | ISEAR | trained preprocessed data on four classifiers and realized that logistic regression outperformed the other methods, that is, SVM, K‐Nearest Neighbor (KNN) and the XG‐Boost with a Precision of 86*%*, recall of 84*%* and an F‐ Score of 85*%* | • Weak context information extraction |
|  |  |  |  | • Recommended the use of a deep learning technique for an improved performance |
| Hasan et al, 2019[49](https://onlinelibrary.wiley.com/doi/10.1002/eng2.12189?fbclid=IwAR2spFkZuRYFGN5sJOoDtlhz1AVbUskUU7QqEy4Jvi0doGZMMPEWPxdoYdU#eng212189-bib-0049) | Hybrid | Tweets | Used SVM, NB and Decision Tree to detect text emotions both online and offline. Obtained an accuracy of 90% | • loose semantic feature extraction |
| Ghanbari‐Adiviand Mosleh, 2019[57](https://onlinelibrary.wiley.com/doi/10.1002/eng2.12189?fbclid=IwAR2spFkZuRYFGN5sJOoDtlhz1AVbUskUU7QqEy4Jvi0doGZMMPEWPxdoYdU#eng212189-bib-0057) | Hybrid | ISEAR, OANC, Tweets | Used K‐NN, MLP and Decision Tree together to extract emotions from texts. Each classifier being 500 in number resulting in a total of 1500 classifiers used. Obtained an accuracy of 99.49% | • Proposed architecture is highly complex |
| Singh et al, 2019[75](https://onlinelibrary.wiley.com/doi/10.1002/eng2.12189?fbclid=IwAR2spFkZuRYFGN5sJOoDtlhz1AVbUskUU7QqEy4Jvi0doGZMMPEWPxdoYdU#eng212189-bib-0075) | Machine Learning | ISEAR | Used POS tagger and Chi‐square for semantic and statistical text feature extraction, and SVM for classification | • Ignored relation between features |
| Nida et al, 2019[96](https://onlinelibrary.wiley.com/doi/10.1002/eng2.12189?fbclid=IwAR2spFkZuRYFGN5sJOoDtlhz1AVbUskUU7QqEy4Jvi0doGZMMPEWPxdoYdU#eng212189-bib-0096) | Hybrid | News Headlines | Classified emotions into 6 using the SVM classifier | • Robust classification technique for improving performance |
|  |  |  |  | • Weak context information extraction |
| Mozafari and Tahayori, 2019[66](https://onlinelibrary.wiley.com/doi/10.1002/eng2.12189?fbclid=IwAR2spFkZuRYFGN5sJOoDtlhz1AVbUskUU7QqEy4Jvi0doGZMMPEWPxdoYdU#eng212189-bib-0066) | Machine Learning | ISEAR | Presented a VSM technique to detect seven emotion categories | • Weak contextual information extraction |
| Tzacheva et al, 2019[97](https://onlinelibrary.wiley.com/doi/10.1002/eng2.12189?fbclid=IwAR2spFkZuRYFGN5sJOoDtlhz1AVbUskUU7QqEy4Jvi0doGZMMPEWPxdoYdU#eng212189-bib-0097) | Hybrid | Tweets | Used the NRC emotion lexicon and SVM to extract actionable emotion patterns in Tweets | • Not good for generalization due to the restricted number of emotion classes |
| Cai and Hao, 2018[78](https://onlinelibrary.wiley.com/doi/10.1002/eng2.12189?fbclid=IwAR2spFkZuRYFGN5sJOoDtlhz1AVbUskUU7QqEy4Jvi0doGZMMPEWPxdoYdU#eng212189-bib-0078) | Machine Learning | Weibo microblogging site | BI‐LSTM based on multiview and attention mechanism were 17*%* better in accuracy than BI‐LSTM detection on either emotion or emoji | • high model complexities . |
| Barbieri et al, 2018[82](https://onlinelibrary.wiley.com/doi/10.1002/eng2.12189?fbclid=IwAR2spFkZuRYFGN5sJOoDtlhz1AVbUskUU7QqEy4Jvi0doGZMMPEWPxdoYdU#eng212189-bib-0082) | Machine Learning | SemEval‐2018 Shared Task on Emoji Prediction | Proposed a labelwise attention LSTM mechanism for detecting emotions in emojis | • model did not suite well for frequently occurring emojis. |
| Jayakrishnan et al, 2018[73](https://onlinelibrary.wiley.com/doi/10.1002/eng2.12189?fbclid=IwAR2spFkZuRYFGN5sJOoDtlhz1AVbUskUU7QqEy4Jvi0doGZMMPEWPxdoYdU#eng212189-bib-0073) | Machine Learning | Malayalam novel | classified sentences in the novel into happy, sad, fear, anger and surprise using the SVM and attained and accuracy of 0.94, 0.92, 0.93, 0.90, and 0.90, respectively. | • disregard for semantic information in texts and the |
|  |  |  |  | • failure to regard the context under which words were used in sentences |
| Allouch et al, 2018[10](https://onlinelibrary.wiley.com/doi/10.1002/eng2.12189?fbclid=IwAR2spFkZuRYFGN5sJOoDtlhz1AVbUskUU7QqEy4Jvi0doGZMMPEWPxdoYdU#eng212189-bib-0010) | Machine Learning | Built dataset from interviews, forums and article comments | SVM obtained a recall of 80% precision of more than 75%, and the Multilayer Neural Network and the Tree Bagger, with recall and precision of more than 75%. | • No semantic representation in model |
|  |  |  |  | • Limited amount of data affected model's performance |
| Tripto and Ali, 2018[76](https://onlinelibrary.wiley.com/doi/10.1002/eng2.12189?fbclid=IwAR2spFkZuRYFGN5sJOoDtlhz1AVbUskUU7QqEy4Jvi0doGZMMPEWPxdoYdU#eng212189-bib-0076) | Machine Learning | YouTube Comments | Emotion Classification Accuracy: 59.2 % Multiclass sentiment labels: 65.97% and 54.24%. | • Satisfactory accuracy results |
| Almanie et al, 2018[62](https://onlinelibrary.wiley.com/doi/10.1002/eng2.12189?fbclid=IwAR2spFkZuRYFGN5sJOoDtlhz1AVbUskUU7QqEy4Jvi0doGZMMPEWPxdoYdU#eng212189-bib-0062) | Rule Based | Tweets | Detected 5 categories of emotions in a real‐time application | • Few number of emotion categories represented results in low generalization ability of proposed system |
|  |  |  |  | • Weak semantic information extraction |
| • Limited vocabulary in lexicon may result in inaccurate classifications |  |  |  |  |
| Ramalingam et al, 2018[38](https://onlinelibrary.wiley.com/doi/10.1002/eng2.12189?fbclid=IwAR2spFkZuRYFGN5sJOoDtlhz1AVbUskUU7QqEy4Jvi0doGZMMPEWPxdoYdU#eng212189-bib-0038) | Hybrid | Twitter Graph API for tweet extraction | Satisfactory performance in classifying emotions | • Robust classification technique for improving performance |
| Abdullah et al, 2018[77](https://onlinelibrary.wiley.com/doi/10.1002/eng2.12189?fbclid=IwAR2spFkZuRYFGN5sJOoDtlhz1AVbUskUU7QqEy4Jvi0doGZMMPEWPxdoYdU#eng212189-bib-0077) | Machine Learning | Arabic Tweets | Compared the performance of Feed‐Forward Neural Network to CNN‐LSTM networks. Obtained an accuracy of 40% and 60% respectively | • Relatively small quantum of data |
|  |  |  |  | • Few number of hidden layers |
| Badugu and Suhasini, 2017[59](https://onlinelibrary.wiley.com/doi/10.1002/eng2.12189?fbclid=IwAR2spFkZuRYFGN5sJOoDtlhz1AVbUskUU7QqEy4Jvi0doGZMMPEWPxdoYdU#eng212189-bib-0059) | Rule Based | Twitter Messages | The system detected 4 emotions with 85% accuracy | • Low generalization ability of proposed system due to limited categories |
|  |  |  |  | • Weak contextual information extraction |
|  |  |  |  | • Overlooked emotions in emoticons |
| Rabeya et al, 2017[63](https://onlinelibrary.wiley.com/doi/10.1002/eng2.12189?fbclid=IwAR2spFkZuRYFGN5sJOoDtlhz1AVbUskUU7QqEy4Jvi0doGZMMPEWPxdoYdU#eng212189-bib-0063) | Rule Based | Facebook status', Newslines, Text books, Direct speeches | Detected 2 emotions with 77.16% accuracy | • Few number of categories. |
|  |  |  |  | • Limited words available in lexicon used |
|  |  |  |  | • Disregard for the context under which words were used . |
| Jain et al, 2017[95](https://onlinelibrary.wiley.com/doi/10.1002/eng2.12189?fbclid=IwAR2spFkZuRYFGN5sJOoDtlhz1AVbUskUU7QqEy4Jvi0doGZMMPEWPxdoYdU#eng212189-bib-0095) | Hybrid | RSS feeds from Twitter | 72.81% accuracy rate obtained | • Overlooked contextual information |
| Kušen et al, 2017[64](https://onlinelibrary.wiley.com/doi/10.1002/eng2.12189?fbclid=IwAR2spFkZuRYFGN5sJOoDtlhz1AVbUskUU7QqEy4Jvi0doGZMMPEWPxdoYdU#eng212189-bib-0064) | Rule Based | Facebook and Twitter Messages | Word‐Emotion Lexicons are capable of identifying emotion categories and Valence. NRC lexicon recorded the highest values for Precision, Recall and F‐measure. EmoSentiNet gave the least values. | • Limited words available in lexicon used. |
|  |  |  |  | • High false positives |
|  |  |  |  | • Few categories represented can result in misclassifications |
| • LeCompte and Chen, 2017[94](https://onlinelibrary.wiley.com/doi/10.1002/eng2.12189?fbclid=IwAR2spFkZuRYFGN5sJOoDtlhz1AVbUskUU7QqEy4Jvi0doGZMMPEWPxdoYdU#eng212189-bib-0094) | Hybrid | Tweets | Analysed the effects of including emojis in detecting emotions from texts | • Robust classification technique for improving performance |
| Mashal and Asnani, 2017[72](https://onlinelibrary.wiley.com/doi/10.1002/eng2.12189?fbclid=IwAR2spFkZuRYFGN5sJOoDtlhz1AVbUskUU7QqEy4Jvi0doGZMMPEWPxdoYdU#eng212189-bib-0072) | Machine Learning | Tweets | detected the intensity of four categories of emotions namely happy, sad, angry and fear in 140 tweets. | • Relatively small quantum of data |
|  |  |  |  | • Restricted number of emotion categories |
|  |  |  |  | • Disregarded for strength of adjectives |
| Grover and Verma, 2016[92](https://onlinelibrary.wiley.com/doi/10.1002/eng2.12189?fbclid=IwAR2spFkZuRYFGN5sJOoDtlhz1AVbUskUU7QqEy4Jvi0doGZMMPEWPxdoYdU#eng212189-bib-0092) | Hybrid | HC Corpora | Model capable of detecting emotions in texts | • Satisfactory predictions |
| Perikos and Hatzilygeroudis, 2016[93](https://onlinelibrary.wiley.com/doi/10.1002/eng2.12189?fbclid=IwAR2spFkZuRYFGN5sJOoDtlhz1AVbUskUU7QqEy4Jvi0doGZMMPEWPxdoYdU#eng212189-bib-0093) | Hybrid | ISEAR and the Affective texts Datasets | Satisfactory performance in classifying emotions | • Non‐robust classification technique affected system performance |
| Wikarsa and Thahir, 2015[71](https://onlinelibrary.wiley.com/doi/10.1002/eng2.12189?fbclid=IwAR2spFkZuRYFGN5sJOoDtlhz1AVbUskUU7QqEy4Jvi0doGZMMPEWPxdoYdU#eng212189-bib-0071) | Machine Learning | Twitter Messages | Detected 5 categories with 83% accuracy | • Relatively small number of tweets |
|  |  |  |  | • Weak context information extraction |
|  |  |  |  | • A more robust ML algorithm can improve performance |
| Lee et al, 2015[69](https://onlinelibrary.wiley.com/doi/10.1002/eng2.12189?fbclid=IwAR2spFkZuRYFGN5sJOoDtlhz1AVbUskUU7QqEy4Jvi0doGZMMPEWPxdoYdU#eng212189-bib-0069) | Machine Learning | Weibo texts | Use Multiview learning to detect emotions.Obtained F1 score of 0.486 | • Word to word translation can result in fuzzy emotion classification |
|  |  |  |  | • Disregard for Negations may result in misclassifications |

[[25]](#footnote-25)

Annex 2: Excel file Tweaked Bert Model

**Other References/Links:**

**Persuasive Design**

* <https://blog.crobox.com/article/marketing-triggers-fbm?utm_campaign=Submission&utm_medium=Community&utm_source=GrowthHackers.com>
* <https://medium.com/crobox/understanding-and-influencing-online-behavior-with-persuasive-design-b5d6e295436a>
* <https://www.uxbooth.com/articles/the-power-and-danger-of-persuasive-design/>
* <https://uxdesign.cc/persuasive-design-nudging-users-in-the-right-direction-5af4a6f8c06f>
* <https://uxdesign.cc/the-ethical-boundaries-of-persuasive-design-c0b040906386>
* <https://www.interaction-design.org/literature/topics/persuasive-design>
* <https://uxdesign.cc/persuasive-design-nudging-users-in-the-right-direction-5af4a6f8c06f>
* <https://instapage.com/blog/bj-fogg-behavior-model>
* <https://endregion.ir/uploads/weblog/persuasive_technology_ref/Fogg%20Behavior%20Model.pdf>

**Design Trends**

* <https://uxplanet.org/flat-design-history-benefits-and-practice-c2b092955f14>
* <https://www.printi.com.br/blog/6-coisas-que-voce-deve-saber-sobre-flat-design>
* <https://blog.revendakwg.com.br/inspiracao-design/flat-design-o-que/>
* <https://www.interaction-design.org/literature/topics/material-design>
* <https://www.treinaweb.com.br/blog/o-que-e-material-design>
* <https://blog.cronapp.io/material-design/>
* <https://www.eclipsemediasolutions.com/blog/what-material-design>
* <https://xd.adobe.com/ideas/principles/web-design/flat-vs-material-skeuomorphic-examples/>
* <https://www.gowtham.design/blog/material-vs-flat>
* <https://99designs.pt/blog/trends/skeuomorphism-flat-design-material-design/>
* <https://blog.prototypr.io/googles-material-design-vs-apple-s-flat-design-which-is-better-16a94aed6e2>
* <https://gapsystudio.medium.com/flat-design-vs-material-design-what-is-the-difference-between-them-5856e102bc66>
* <https://www.appventurez.com/blog/material-design-vs-flat-design/>
* <https://www.sktthemes.org/website/flat-design-vs-material-design/>
* <https://uxplanet.org/flat-design-vs-material-design-whats-your-flavor-43a27c295f62>

**Interfaces for the elderly:**

* <https://www.toptal.com/designers/ui/ui-design-for-older-adults>
* <https://www.uxmatters.com/mt/archives/2017/04/designing-user-interfaces-for-an-aging-population.php>
* <https://uxplanet.org/ux-accessibility-for-elderly-12-principles-9708289b6f78>
* <https://www.researchgate.net/publication/321757825_Designing_User_Interfaces_for_the_Elderly_A_Systematic_Literature_Review>
* <https://uxplanet.org/ux-study-designing-for-older-people-6c67575d9c2f>
* <https://f5-studio.com/articles/age-friendly-ui-ux-design-thinking-for-senior-citizens/>

**NLP ML Models/Datasets:**

* <https://github.com/lukasgarbas/nlp-text-emotion>
* <https://github.com/Ankit152/Emotions-Detection-NLP>
* <https://medium.com/neuronio/from-sentiment-analysis-to-emotion-recognition-a-nlp-story-bcc9d6ff61ae>
* [Learning local and global contexts using a convolutional recurrent network model for relation classification in biomedical text (aclweb.org)](https://www.aclweb.org/anthology/K17-1032.pdf)
* <https://www.researchgate.net/publication/340804581_Combining_Local_and_Global_Features_Into_a_Siamese_Network_for_Sentence_Similarity>
* <https://github.com/atulapra/Emotion-detection>
* <https://www.kaggle.com/praveengovi/emotions-dataset-for-nlp>
* <https://paperswithcode.com/paper/dialoguernn-an-attentive-rnn-for-emotion#code>
* <https://paperswithcode.com/paper/dialoguegcn-a-graph-convolutional-neural>
* <https://paperswithcode.com/paper/multimodal-speech-emotion-recognition-using>
* <https://github.com/maelfabien/Multimodal-Emotion-Recognition>
* <https://github.com/emorynlp/emotion-detection>
* <https://github.com/Harsh24893/EmotionRecognition/blob/master/README.md>
* <https://github.com/karankishinani/Emotion-detection-from-text-using-PyTorch-and-Federated-Learning>
* <https://towardsdatascience.com/natural-language-processing-pipeline-decoded-f97a4da5dbb7>

<https://medium.com/@suneelpatel.in/nlp-pipeline-building-an-nlp-pipeline-step-by-step-7f0576e11d08>

1. Borod JC. The Neuropsychology of Emotion. Oxford, UK: Oxford University Press; 2000. [↑](#footnote-ref-1)
2. <https://www.paulekman.com/blog/darwins-claim-universals-facial-expression-challenged/> [↑](#footnote-ref-2)
3. <https://www.nature.com/articles/d41586-020-00507-5> [↑](#footnote-ref-3)
4. Plutchik R. A general psychoevolutionary theory of emotion. Amsterdam, Netherlands: Elsevier; 1980 (pp. 3–33). [↑](#footnote-ref-4)
5. <https://www.6seconds.org/2020/08/11/plutchik-wheel-emotions/> [↑](#footnote-ref-5)
6. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4243519/#:~:text=The%20model%20proposed%20by%20Ortony%2C%20Clore%2C%20and%20Collins%20(commonly,and%20those%20focused%20on%20objects> [↑](#footnote-ref-6)
7. What does an artistic scientist create? | Shunji Mitsuyoshi | TEDxTokyo. 4 August 2015 <https://www.youtube.com/watch?v=oOsst8jHMxQ> [↑](#footnote-ref-7)
8. The Economic Time. SoftBank’s struggles with Pepper keep Son’s robot dreams on hold. <https://economictimes.indiatimes.com/small-biz/security-tech/technology/softbanks-struggles-with-pepper-keep-sons-robot-dreams-on-hold/articleshow/55108707.cms?utm_source=contentofinterest&utm_medium=text&utm_campaign=cppst> [↑](#footnote-ref-8)
9. <https://www.theverge.com/2019/7/25/8929793/emotion-recognition-analysis-ai-machine-learning-facial-expression-review> [↑](#footnote-ref-9)
10. <https://jov.arvojournals.org/article.aspx?articleid=2138723> [↑](#footnote-ref-10)
11. <https://onlinelibrary.wiley.com/doi/10.1002/eng2.12189?fbclid=IwAR2spFkZuRYFGN5sJOoDtlhz1AVbUskUU7QqEy4Jvi0doGZMMPEWPxdoYdU#eng212189-bib-0011> [↑](#footnote-ref-11)
12. Jacob Devlin & Ming-Wei Chang. [Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing](http://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html). Friday, November 2, 2018. <https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html?fbclid=IwAR1ohUYvopIZAmWUK2qpkn3k29wTfGsgrlCR-p29R2vuO22-f2hUsDsHqKk> [↑](#footnote-ref-12)
13. [jacobdevlin-google](https://github.com/google-research/bert/commits?author=jacobdevlin-google). [google-research](https://github.com/google-research)/[bert](https://github.com/google-research/bert). <https://github.com/google-research/bert> [↑](#footnote-ref-13)
14. For more information on how the transformer works we refer to <http://jalammar.github.io/illustrated-transformer/> [↑](#footnote-ref-14)
15. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Google AI Language {jacobdevlin,mingweichang,kentonl,kristout}. <https://arxiv.org/pdf/1810.04805.pdf> [↑](#footnote-ref-15)
16. Chris McCormick. BERT Research. <https://www.youtube.com/watch?v=FKlPCK1uFrc&t=3s> [↑](#footnote-ref-16)
17. [Ask Your Neurons: A Neural-Based Approach to Answering Questions About Images (cv-foundation.org)](https://www.cv-foundation.org/openaccess/content_iccv_2015/papers/Malinowski_Ask_Your_Neurons_ICCV_2015_paper.pdf) [↑](#footnote-ref-17)
18. <https://link.springer.com/article/10.1007/s10115-020-01449-0/tables/3> [↑](#footnote-ref-18)
19. <https://onlinelibrary.wiley.com/doi/10.1002/eng2.12189?fbclid=IwAR2spFkZuRYFGN5sJOoDtlhz1AVbUskUU7QqEy4Jvi0doGZMMPEWPxdoYdU#eng212189-bib-0011> [↑](#footnote-ref-19)
20. <https://github.com/lukasgarbas/nlp-text-emotion> [↑](#footnote-ref-20)
21. <https://huggingface.co/transformers/preprocessing.html> [↑](#footnote-ref-21)
22. <https://ruder.io/deep-learning-optimization-2017/index.html#warmrestarts> [↑](#footnote-ref-22)
23. <https://onlinelibrary.wiley.com/doi/10.1002/eng2.12189?fbclid=IwAR2spFkZuRYFGN5sJOoDtlhz1AVbUskUU7QqEy4Jvi0doGZMMPEWPxdoYdU#eng212189-bib-0011> [↑](#footnote-ref-23)
24. <https://onlinelibrary.wiley.com/doi/10.1002/eng2.12189?fbclid=IwAR2spFkZuRYFGN5sJOoDtlhz1AVbUskUU7QqEy4Jvi0doGZMMPEWPxdoYdU#eng212189-bib-0011> [↑](#footnote-ref-24)
25. <https://link.springer.com/article/10.1007/s10115-020-01449-0/tables/3> [↑](#footnote-ref-25)