An Insight into Emotion Recognition Applications: An Evaluation and Comparison Between Amazon's Rekognition and IBM's Watson Natural Language Understanding

[redacted]

18/01/23

I. INTRODUCTION

A. *Purpose/Aim*

This paper provides an assessment and discussion regarding emotion recognition applications (ERAs) within affective computing, by providing an analysis and comparison between two interactive and functional ERAs. These are monomodal ERAs, which make use of different data types, to see how the design of users-AI interaction differs between these systems when predicting emotion.

The ERAs selected include Amazon's Rekognition [4, 5] and IBM's Watson Natural Language Understanding (WNLU) [6, 7] interactive online demos. To assess these ERAs, a heuristic and empirical evaluation were carried out. Throughout this paper, the key concepts of affective computing will be explained, along with the ERAs and their selection process, the heuristic and empirical evaluation of the ERAs with an in depth analysis and a concluding discussion regarding ERAs.

B. *Key Concepts, Methods and Applications*

An affective definition of emotion is one which has an emphasis on feelings of excitement and pleasure [20]. Emotions have the power to drive us as humans and are related to a person's interaction with the environment and relevance to their life; these are observable and can be analysed within a person by assessing their body language. Emotions are distinct from feelings, sentiment and mood. They are commonly represented in either a discrete manner - describing every emotion from a set of basic emotions, or a dimensional manner – with emotions within a 2D or 3D space.

Affective computing is any form of computation related to emotional states, including computing which can bring about or influence certain emotions of a user [8] through computer vision. The recognition and simulation of voice and body language is particularly common [9] within this rapidly growing field, in addition to processing emotions to provide a particular response. Affective computing also has links to cognitive neuroscience, psychology, philosophy, linguistics and AI which is commonly applied within affective computing to model emotion [10, 23].

There are a diverse range of applications which make use of affective computing, such as affect-based diagnosis in healthcare, video games or educational tools for teaching and wellbeing. These technologies can sense, respond and learn with you [22].The systems either follow a form of one-way communication, where users affect computers (an affect monitor) or computers affect users (an affect simulator), or two-way communication where users and computer influence each other in a cycle. These comprise of affective systems - where there is an emotional form of interaction and the user may be influenced - and affect-aware systems - which also recognise the emotional state of a user [21].

There are many emotion recognition algorithms which make use of machine learning, such as for text analysis, visual information, body movements, voice signals, the usage of input devices and physiological measurements. These differ in the number of input channels, features and methods of data extraction, feature selection and classification. Video, images and text are common as inputs. ERAs may be monomodal and focus on one data type or may be multimodal and focus on many inputs. The output and interface of affective systems must also be considered for the user experience (UX). This can be defined as the perceptions of a user of a product/service that results in their use or anticipated use of that product/service [14].

An example of an ERA is 'AffdexMe' by Affectiva, which is a monomodal ERA that analyses facial expressions with real-time video input, using a discrete representation of emotions. Further examples of ERAs found during the search process can be seen below.

II. METHOD

A. *Search Process*

To identify popular ERAs, a search for "emotion recognition apps" was simply made through Google and apps were also sought for on sourceforge.net [24] which appeared in the search. This process was chosen as it is the most simple and reliable way to source apps using emotion recognition, as search engines rank results in terms of relevance.

Popular ERAs found included those on app stores, such as 'Emotimeter' [25] on the Google store, which uses deep learning to detect emotions from facial expressions. ERAs were also found in the form of SDK/APIs which could be installed to be included in the development of affective learning products, such as 'Behavioural Signals' [26] which uses speech data to match customers to the most appropriate agent that can handle their call.

B. *Description of ERAs*

The ERAs chosen include Amazon Rekognition and IBM Watson Natural Language Understanding [4, 6].

Rekognition performs facial analysis in images and video, allowing automatic emotional analysis through the API. It also serves as a tool to label or identify a variety of objects, places and activities and one which can detect, analyse and compare these.

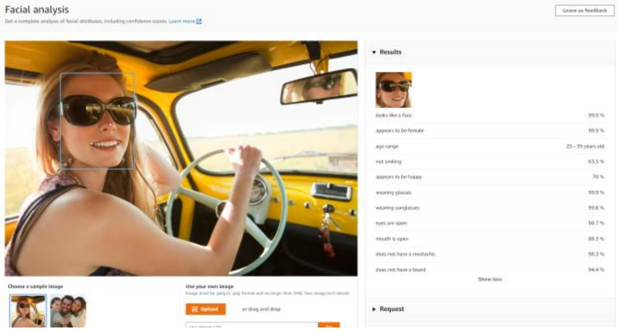
This model is trained through deep learning and the daily analysis of billions of images and videos.

WNLU extracts meta-data from written text, allowing automatic emotional and sentiment analysis through its API. This is completed through natural language processing and the use of deep learning models. WNLU is also able to extract a variety of features from text, such as keywords, and analyse linguistics, such as identifying the function of each word within a sentence.

Both analysis tools have online demonstrations [5, 7] – this study will specifically assess these two online applications. The Rekognition and WNLU APIs were installed and analysed also, ensuring that the output was consistent when comparing the direct use of the API and the application outputs that were produced.

The Rekognition 'Facial Analysis' demo works solely with images. The application contains two sample images which can be analysed, or the user can upload or 'drag and drop' or provide an image URL. This system provides a variety of statistics, including whether there appears to be a face, the assumed gender of the face in the image, their approximate age, whether the person is happy or smiling, whether they are wearing glasses or sunglasses, whether they have facial hair and whether their eyes or mouth are open. The 'appears to be happy' statistic will be focussed upon.

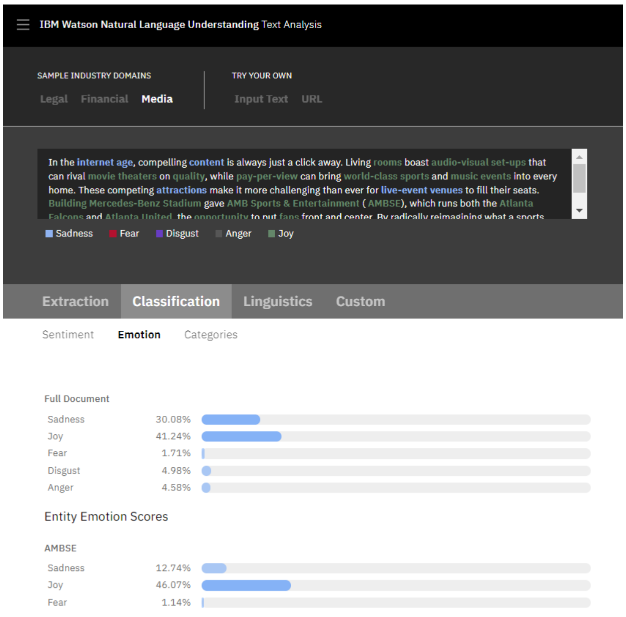
The WNLU application contains a 'legal', 'financial' and 'media' examples and allows text, raw HTML or a URL to be provided. The application is also able to extract 'entities', 'keywords', 'concepts', 'relations', classify the 'sentiment', 'emotion' and 'categories' a sentence belongs to, and analyse the language usage including the 'syntax' and 'semantic' roles. In this study, the emotion classification will be focussed upon. Certain words are highlighted by the system, depending on the results that are provided.



Results

looks like a face	99.9 %
appears to be female	99.9 %
age range	25 - 35 years old
not smiling	63.5 %
appears to be happy	70 %
wearing glasses	99.9 %
wearing sunglasses	99.8 %
eyes are open	98.7 %
mouth is open	88.5 %
does not have a mustache	98.3 %
does not have a beard	94.4 %

Fig 1. The Amazon Rekognition application / demo, with a sample image as input.



In the internet age, compelling content is always just a click away. Living rooms boast audio-visual set-ups that can rival movie theaters on quality, while pay-per-view can bring world-class sports and music events into every home. These competing attractions make it more challenging than ever for live-event venues to fill their seats. Building Mercedes-Benz Stadium gave AMB Sports & Entertainment (AMBSE), which runs both the Atlanta Falcons and Atlanta United, the opportunity to put fans front and center. By radically reimagining what a sports

🟦 Sadness 🟥 Fear 🟪 Disgust ⬛ Anger 🟩 Joy

Full Document

Sadness	30.08%
Joy	41.24%
Fear	1.71%
Disgust	4.98%
Anger	4.58%

Entity Emotion Scores

AMBSE

Sadness	12.74%
Joy	46.07%
Fear	1.14%

Fig 2. The IBM Watson Natural Language Understanding application / demo, with sample 'Media' text as input.

C. *Justification of ERA Choices*

These ERAs were chosen, since they are both monomodal ERAs which make use of different data types, which aligns with the aims of this study, with Rekognition using face images and WNLU making use of inputted text.

Additionally, both applications were free and had available API which was explored through Python. This was not applicable for many ERAs available on the internet, such as 'Emotimeter' and 'Behavioural Signals' – differentiating Rekognition and WNLU from these other apps.

IBM and Amazon are trusted world-leading organisations in the technologies and innovations they bring to the world, which work to a professional standard. Consequently, as Rekognition and WNLU were created by a reliable company, they were chosen to be evaluated over other ERAs. Moreover, Rekognition and WNLU are widely used when compared to some other ERAs and any improvements to these applications can assist all the other applications which integrate these APIs.

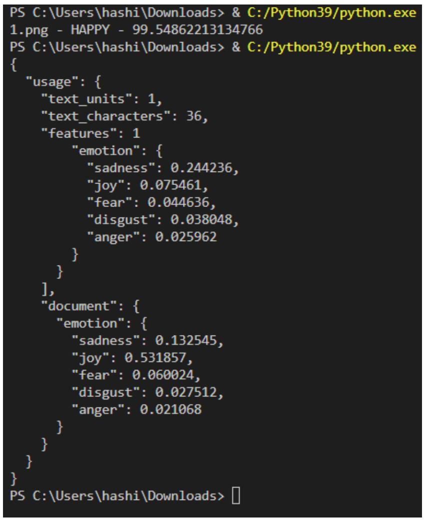
These applications are easily accessible online and hence are easily testable for a variety of users, assisting the process of a heuristic and empirical evaluation.

D. *Installation Process*

Although both ERAs were accessible through the web browser, their SDKs were also installed and analysed offline as discussed previously.

To setup the SDK and install the API for Amazon's Rekognition, a user account had to be created with Amazon Web Services (AWS). An IAM (Identity and Access Management) account was then created through the AWS service allowing JSON requests to be made. To use Rekognition, boto3 was installed through the command line. In Python, using the code example implementations from AWS [1] the application could be used. After checking the security details for the IAM user that was created, references to the AWS access key ID, secret access key and region name were made, allowing the code to work as intended.

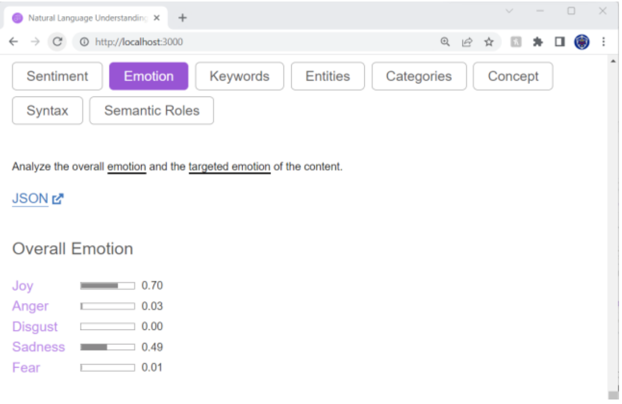
Similarly for IBM's WNLU, a user account and IAM account had to be created, after building the WNLU service in the IBM cloud service. Once again, code implementations from IBM [2] were applied in Python. References to the IAM ID and region service URL were made within this code, ensuring it connected to the service to run as intended.



```
PS C:\Users\hashi\Downloads> & C:/Python39/python.exe
1.png - HAPPY - 99.54862213134766
PS C:\Users\hashi\Downloads> & C:/Python39/python.exe
{
  "usage": {
    "text_units": 1,
    "text_characters": 36,
    "features": 1
        "emotion": {
          "sadness": 0.244236,
          "joy": 0.075461,
          "fear": 0.044636,
          "disgust": 0.038048,
          "anger": 0.025962
        }
      }
    ],
    "document": {
      "emotion": {
        "sadness": 0.132545,
        "joy": 0.531857,
        "fear": 0.060024,
        "disgust": 0.027512,
        "anger": 0.021068
      }
    }
  }
}
PS C:\Users\hashi\Downloads> 
```

Fig 14. Example outputs through the API, for Rekognition and WNLU respectively.

A version of WNLU was also tested, which runs offline and is launched off the local host. The code from an IBM GitHub repository [3] was simply downloaded and the instructions on the GitHub page were followed.



Analyze the overall emotion and the targeted emotion of the content.

Overall Emotion

Joy 0.70
Anger 0.03
Disgust 0.00
Sadness 0.49
Fear 0.01

Fig 3. The WNLU offline application, showing the output for the custom phrase "I love human AI interaction".

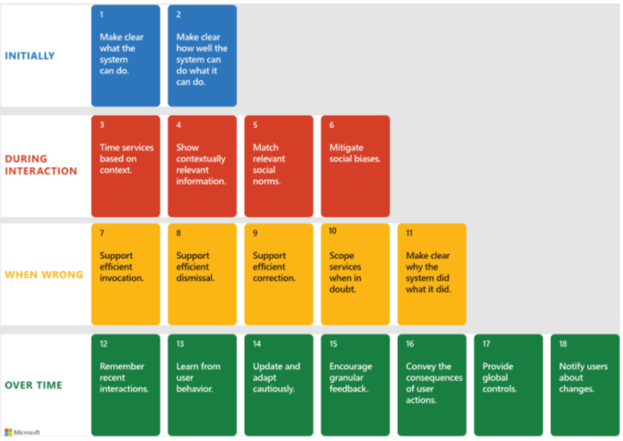
III. EVALUATION

It is important for usability to be high within systems, ensuring that user needs are always considered to create efficient, effective and satisfactory system. These systems should simply contain everything that is necessary and contain features developed that are used as intended. This can help users make full use of a system, ensuring they are in control and their goals are respected. Hence, this study makes use of two usability evaluation methods (UEMs) to assess the two ERAs.

STUDY 1: HEURISTIC EVALUATION

A. *Evaluation Procedure*

Within this study, an analytical UEM in the form of a heuristic evaluation is carried out. Usability problems in a user interface design are quickly sought out through using a set of guidelines/ evaluators to judge and inspect how well usability principles (heuristics) are adhered to [12]. There are a variety of guidelines which can be used, many of which are all derived from Nielsen's ten usability principles [13] for UX design.



INITIALLY
1 Make clear what the system can do.
2 Make clear how well the system can do what it can do.

DURING INTERACTION
3 Time services based on context.
4 Show contextually relevant information.
5 Match relevant social norms.
6 Mitigate social biases.

WHEN WRONG
7 Support efficient invocation.
8 Support efficient dismissal.
9 Support efficient correction.
10 Scope services when in doubt.
11 Make clear why the system did what it did.

OVER TIME
12 Remember recent interactions.
13 Learn from user behavior.
14 Update and adapt cautiously.
15 Encourage granular feedback.
16 Convey the consequences of user actions.
17 Provide global controls.
18 Notify users about changes.

Fig 4. The 18 Microsoft guidelines for human-AI interaction.

The Microsoft heuristic evaluation guidelines [19, 27] were used, as this is commonly used at an industry standard. Even though this evaluation takes longer, the 18 guidelines are able to provide a thorough evaluation, to help find the best way to apply different design principles.

To carry out the heuristic evaluation, each guideline is assessed one by one while using the application interface. The heuristics are split into four groups, including an 'initial phase' with users' assumptions of the application based on the information provided (G1-G2), the experience of the application 'during the interaction' (G3-G6), 'when wrong' (G7-G11) and 'over time' (G12-G18), as demonstrated in *Fig 4*. These phases were assumed in order during the evaluation. It is ensured that a table is updated with the results of each guideline assessment. This process is run through twice and a violation is marked as having occurred if it is reported on either pass.

Severity	Colour
Minor	🟩
Major	🟨
Catastrophe	🟥

Fig 6. The colour scheme for severity rating.

A severity rating, such as one outlined for usability problems by the Nielsen Norman Group [16], is commonly used by evaluators upon the problems they discover. An adapted version of this severity rating will be used, including a minor (low priority issue), major (high priority issue) and catastrophe (must be fixed immediately) ranking for the violations. A guideline, within the table of results, that has been violated will be coloured following the scheme in *Fig 5*.

B. Report of Results

Guideline	Application	Violation	Does Not Apply
G1		✓	
G2		✓	
G3			✓
G4		✓	
G5	✓		
G6	✓		
G7	✓		
G8		✓	
G9		✓	
G10		✓	
G11	✓		
G12		✓	
G13		✓	
G14	✓		
G15		✓	
G16	✓		
G17		✓	
G18	✓		

Fig 6. The heuristic evaluation table for the Amazon Rekognition application / demo.

Applications

G5 – The application accepts a wide range of photos.

G6 – The application can identify a wide range of faces, no matter who the person is or where they are from or what their gender is. The example images are appropriate.

G7 – There is a clear indication of where to input or upload an image, when the user would like to.

G11 – When an image without a face is supplied, the user is notified that no faces are detected, and no statistics appear.

G14 – Updates the statistics appropriately and displays the image when there is a new one. The rest of the system is kept consistent.

G16 – Uploading an image instantaneously updates the application and statistics.

G18 – Notifications for updates to the AI are assumed to be in place, such as the ‘new’ in the other demos available.

Violations

G1 – A short description of the system is explained at the top of the application, however this is not detailed enough. There is although a link to learn more.

G2 – There is a link to a learn more page, however there is no indication of how well the system performs.

G4 – The application shows either ‘appears to be happy/appears not to be happy’ and updates the percentage for the image on the screen. There is however no further information to what has been done.

G8 – It is difficult to solely focus on the facial emotion recognition aspect and filter out all the other statistics.

G9 – There is only an option to re-upload images if there is an error, but not one to report an error or go back to a system state.

G10 – The system does not provide alternative options when an image without a face is supplied, however the user is notified there is no face.

G12 – Previous images must be uploaded again to get their statistics. There is no button to go back to these states.

G13 – Although Rekognition constantly improves from new data, it does not learn user behaviour to provide a personalised response, such as suggest other images if faces are constantly not provided.

G15 – The user cannot provide feedback to the system over what they preferred. Such as no customisability for different emotions.

G17 – There is no customisability feature.

Guideline	Application	Violation	Does Not Apply
G1	✓		
G2		✓	
G3			✓
G4	✓		
G5	✓		
G6	✓		
G7	✓		
G8	✓		
G9		✓	
G10		✓	
G11	✓		
G12		✓	
G13		✓	
G14	✓		
G15		✓	
G16	✓		
G17		✓	
G18		✓	

Fig 7. The heuristic evaluation table for the IBM Watson Natural Language Understanding application / demo.

Applications

G1 – There is an information box at the right of the screen, which explains what WNLU does. There is a link to learn more.

G4 – There is an information box and a link detailing the emotion analysis. Statistics are provided and words are appropriately highlighted.

G5 – The application accepts a wide range of sentences.

G6 – The application can identify a wide range of sentences and does not conform to stereotypes and biases.

G7 – The text can be easily changed, when necessary, through selecting an example or inputting text / a URL.

G8 – The emotion analysis tool can be selected, and all the statistics displayed are relevant to this and the words within the input.

G11 – The text which cannot be analysed returns ‘no results available’.

G14 – The system is kept constant while the statistics and highlighted words appropriately change after the text is edited.

G16 – The system instantly updates when new text is entered.

Violations

G2 – There is no indication as to how well the system performs.

G9 – The text can be adjusted if there is an error, however there is no way to go back to previously entered text or report an error.

G10 – The system does not provide alternative options when the text cannot be analysed, however the user is notified that there are no results available.

G12 – The input text can only be edited. The system does not remember previous inputs and results.

G13 – WNLU constantly improves from new data, however a personalised user experience is not created by learning user patterns.

G15 – There is an option to provide feedback at the right of the screen, however this is not actively encouraged.

G17 – There is a customisation option where the user can define ‘entities’ and ‘relations’, however this is ‘only available for the trained models used in the ‘Sample Industry Domains’.

G18 – There is no evidence for the notification of updates to the AI.

C. *Comparison and Discussion of Results*

Guideline	Shared Application (✓) / Violation (✕)
G1	–
G2	✕
G3	N/A
G4	–
G5	✓
G6	✓
G7	✓
G8	–
G9	✕
G10	✕
G11	✓
G12	✕
G13	✕
G14	✓
G15	✕
G16	✓
G17	✕
G18	–

Fig 8. The shared applications and violations, with ‘✓’ representing a shared application, ‘✕’ representing a shared violation and ‘’ representing an application in one ERA and a violation in the other.

Both demos faced a series of successes and failures after going through the heuristic evaluation – many of which were shared, as seen in Fig 8. 7 applications and 10 violations were found in the Rekognition application, whereas 9 applications and 8 violations were found in the WNLU application, suggesting that the latter has a better interface design for human-AI interaction.

When considering all phases, WNLU performs better. However, when looking at each phase individually, Rekognition completely failed the ‘initial’ phase. WNLU passed all the ‘during interaction’ phase, whereas Rekognition only failed G4 here. Both ERAs had mixed results for the final two phases, with WNLU having one more application (G8) in the ‘when wrong’ phase and Rekognition with one more application (G18) in the ‘over time’ phase.

It must be ensured that the two ERAs take into consideration ‘human control, transparency, explainability, fairness, justice, inclusiveness, sustainability, and education’ [28] in order to meet the heuristics and ensure user needs are sufficiently met, as outlined previously. WNLU does well containing an information box at the side of the screen and an external link to the details regarding the emotional analysis, which ensures that users are completely aware of the functionality of the system. On the other hand, Rekognition contains fewer details, and so these should be added to the application for users to be aware of, such as the suggestion to upload an image. This outlines the main issue with the Rekognition demonstration – that there is not enough detail regarding emotional analysis, including the statistics provided.

There are some things that demonstrate the good UX design of Rekognition and WNLU. Both ERAs can accept a wide range of inputs, and these can be adjusted when necessary. Additionally, both applications can adapt well to changing inputs, including explaining to the user when they have provided an inappropriate input and instantly adjusting the displayed statistics appropriately. On the other hand, both WNLU and Rekognition do not allow users to go back to previous inputs to correct errors, and they do not allow users to report any errors. These functions should be added to ensure that users are less confused and will feel supported when something goes wrong. Finally, patterns of user interactions are not learned by either system – this integration could help make the systems feel more personal for users.

STUDY 2: EMPIRICAL EVALUATION

D. Evaluation Procedure

It is important to carry out a user-based empirical evaluation when evaluating usability and UX. This is defined as the testing of factors through user observations in experiments, to evaluate the software system with a user model [15].

To evaluate usability, most studies focus on the error rate for effectiveness, the task completion time for efficiency and a satisfaction rating with the interface as a satisfaction measure [11]. The system usability scale (SUS) is also often used, which covers these concepts. 15 people are ideally required [17] to ensure that as much of the system is tested as possible.

According to the Nielsen Normal group, five participants are sufficient to provide the best results [29] when carrying out usability testing to reduce costs and test as much of an application as possible. As a result, this study made use of convenience sampling of five participants, who all had mixed attributes and backgrounds and were not incentivised to agree with an interviewer.

Task Number	Task for Rekognition	Time to Complete
1	Find out how happy the face is in the example image.	30 seconds
2	Select the second sample image and find out how happy the little girl is.	45 seconds
3	Upload the pre-saved image of a face (‘Michelle Yeoh’ already saved on the laptop) and find out how happy it is.	1 minute
4	Paste an image URL (already accessible) of a face and find out how happy it is.	45 seconds

Task Number	Task for WNLU	Time to Complete
5	Select the ‘media’ example text and analyse it and locate the emotion analysis.	30 seconds
6	Select the ‘financial’ example text and see the percentage of joy for the word ‘quarter’.	45 seconds
7	Input the text ‘I love to eat pineapple pizza on Mondays’ and find out the percentage of sadness in the sentence.	1 minute
8	Input the URL (already accessible) and find out the percentage of disgust	45 seconds

	for the word ‘bazinga’.	

Fig 9. The tasks to be carried out by users within a certain consistent timeframe.

To carry out the empirical evaluation, each participant was seated one-on-one with the interviewer. The process that follows was carried out for each ERA in a random order. Firstly, the participants were required to complete frequently performed tasks/goals, as seen within *Fig 9*, within a set timeframe. These were realistic tasks, considering the intentions of the application. It was ensured that the think-aloud protocol [18] was used, allowing users to continuously verbalise and elaborate upon their thoughts and to find out why they were carrying out certain tasks. This was completed concurrently – even though this could have interfered with the participants’ actions – as recalling can be incorrect and the participants’ answers could be influenced by their experience in later tasks (as experience may improve when considering a UX curve). Data of these responses were recorded on paper in a bullet point form during the conversation.

Additionally, it was ensured that participants completed a trust scale [30] and SUS questionnaires [32] once the interview was over. This made certain that data was collected of user expectations, before interacting with either ERA, during using, and afterwards to reflect on their experience. Users could have drawn out a UX curve [33] or the SAM or AttrakDiff [31] test could have been carried out if there was more time, however they were not within this study due to the small scale nature of the tests carried out. This is the same for psychophysiological measurements, such as an EEG, which are common for UX evaluation, however they were not used here.

The initial purpose of the eight think aloud tasks were used to measure both the intuitiveness of the application and qualitatively identify the areas where the user went wrong. Similarly, the SUS questionnaire (see [32] for the ten questions) was used to measure the personal usability for each participant. The ten trust scale questions (see [30] for the questions) were used to measure trustability. Both the trust and SUS questions range from 1-5 (from strongly disagree to strongly agree).

E. Report of Results

The results were condensed into a small set of tables.

Task Number	Percentage Completed within the Time
1	80%
2	100%
3	80%
4	100%
5	80%
6	100%
7	100%
8	100%

Fig 10. The percentage of tasks completed within the allocated time by participants. See Fig 9 for the tasks.

Participant Number	Comments for Rekognition	Comments for WNLU
1	Unaware of the ‘see more’ button and of the function of the application.	Found it difficult to initially locate the emotion analysis.
2	No issues.	Thought the system was appealing.
3	Had difficulty in uploading the saved image.	Found it difficult to navigate.
4	No issues.	Found it very intuitive.
5	Did not like the system design.	No issues.

Fig 11. Notable comments from participants during the study for each ERA.

Trust Question Number	Average Score for Rekognition	Average Score for WNLU
1	4.0	4.6
2	4.4	4.4
3	3.4	4.6
4	4.0	4.4
5	5.0	5.0
6	1.6	1.6
7	5.0	5.0
8	4.6	4.8

Fig 12. The average trust questionnaire results for each ERA.

SUS Question Number	Average Score for Rekognition	Average Score for WNLU
1	4.0	4.6
2	1.8	1.4
3	4.6	4.8
4	1.8	1.8
5	4.0	4.6
6	1.6	1.4
7	4.6	4.4
8	2.0	1.4
9	4.0	4.4
10	1.4	1.8
Total Score	32.6*2.5 = 81.5	35*2.5 = 87.5

Fig 13. The average SUS questionnaire scores and total scores, using the SUS formula.

F. *Comparison and Discussion of Results*

Based on the interview tasks completed in the empirical evaluation, 90% of the Rekognition task were completed within the appropriate timeframe, whereas 95% of tasks were completed with WNLU. Tasks 1-4 were approximately set to be equally as difficult as tasks 5-8 respectively. Hence, this may indicate that the human-AI design of WNLU is more user friendly in its current state. Moreover, the comments noted down included one more negative for Rekognition than for WNLU. When comparing the trust question scores, the WNLU scores are slightly favoured as they are either equal or lean towards the desired result in almost every question. The largest differences were found in the reliability of the ERAs, as seen in question 3 – the participants found WNLU to be more reliable, which would be due to the perceived robustness of each application. On the other hand, the users did find both systems to be equally efficient and produce results quickly, as seen in question 5. Finally, the SUS score for WNLU was higher than that of Rekognition. Once again, the results one each question were similar, however the WNLU scores were slightly better or the same. One clear difference is seen in question 1, where users would be more favoured to use WNLU than Rekognition.

When considering the interview tasks and the SUS results, both systems could make their emotion data statistics more easily accessible – as indicated by tasks 1, 3 and 5. It is important for ERAs to have high ease of access [34] to navigate a system, ensuring that users can make use of all the system's function – priority should be put towards the goal of emotion recognition. The trustability of the Rekognition application could be improved by increasing the explainability of the system [35], as it is important for users to know what the system has done – this was found to be more apparent for the WNLU application than Rekognition.

IV. DISCUSSION

A. *Strength, Weaknesses and Socioeconomic Impacts*

From this study, it is clear some ERAs perform better, as demonstrated by WNLU outperforming the Rekognition application in both the heuristic and empirical evaluations. Weaknesses between these ERAs include the need to increase functionality for the user when the system goes wrong and to improve the ease of access and explainability of the system and AI.

The main strength of ERAs includes the wide positive impact they can have within society regarding health and wellbeing, such as the ability to assist the mental state of callers, creating a kinder world. Due to this, the size of the market is set to double by 2027 [37], as new products are integrated and more interest is generated. On the other hand, it is still difficult to validate the accuracy of ERA prediction. Since there is a chance of ERAs being wrong this prevents their widespread use – as it would be unethical to put all faith in these systems.

B. *Ethical Implications*

ERAs must conform to AI ethics norms. The use of ERAs may have impacts in a variety of areas, such as privacy where data must be collected with consent. There may also be issues with the ERAs themselves, such as bias, due to the data the ERA was trained on. There are also issues related to morality, power, ownership, environmental impact, machine dependency and perceived consciousness [11], which all become larger issues as ERAs gain popularity in our world.

C. *Design Enhancements*

As discussed previously, the ERAs previously evaluated should focus more on increasing the functionality when the system goes wrong, as well as the explainability and ease of access of the system. The focus should also be on allowing the user access to the emotional analysis aspect.

In general, the design of ERAs will focus more on multimodal systems, which use more than one data type, to better predict emotion. Similarly, the use of contextual information in ERAs [37] is currently rare, such as using a user's location, although this could provide more accurate results. Also, when training the emotion recognition models, more diverse data could be used, or emotions could be defined in a dimensional rather than a discrete manner. Finally, psychophysiological devices could be reduced in size to be easily transportable to predict emotions in a variety of situations.

V. CONCLUSION

Within this study, affective computing and emotions within this field were initially defined; the variety of applications within affective computing, as well as examples, were explored. The process of finding appropriate ERAs and installing their APIs was explained. A heuristic evaluation (using the 18 Microsoft guidelines) and an empirical evaluation (by interviewing participants and suppling them the SUS and trust tests) were carried out on a couple of interactive monomodal ERAs – the textual-based WNLU and the facial analysis Rekognition. It was concluded that WNLU performed better due to a more clear and trustworthy system. Finally, a discussion about the strengths, weaknesses and future of ERAs was explained, which appears to focus on ethics and systems taking in more data as input.

REFERENCES

[1] *Amazon rekognition examples using SDK for python (boto3), Amazon*. Available at: https://docs.aws.amazon.com/code-library/latest/ug/python_3_rekognition_code_examples.html (Accessed: January 19, 2023).

[2] *Natural language understanding - IBM cloud API docs, IBM*. Available at: https://cloud.ibm.com/apidocs/natural-language-understanding?code=python (Accessed: January 19, 2023).

[3] *Demo code for the natural language understanding service, GitHub*. Available at: https://github.com/watson-developer-cloud/natural-language-understanding-nodejs (Accessed: January 19, 2023).

[4] *Amazon Rekognition, Amazon*. Available at: https://aws.amazon.com/rekognition/ (Accessed: January 19, 2023).

[5] *Facial Analysis, Amazon*. Available at: https://eu-west-2.console.aws.amazon.com/rekognition/home?region=eu-west-2#/face-detection (Accessed: January 19, 2023).

[6] *IBM Watson Natural language understanding – overview, IBM*. Available at: https://www.ibm.com/uk-en/cloud/watson-natural-language-understanding (Accessed: January 19, 2023).

[7] *IBM Watson Natural Language Understanding Text Analysis, DTE NLU Demo*. Available at: https://www.ibm.com/demos/live/natural-language-understanding/self-service/home (Accessed: January 19, 2023).

[8] Picard, R. (1995) *Affective Computing*. Avaliable at: https://affect.media.mit.edu/pdfs/95.picard.pdf (Accessed: January 19, 2023).

[9] Tao, J. and Tan, T. *Affective Computing: A Review*. Available at: https://www.researchgate.net/publication/220270285_Affective_Computing_A_Review (Accessed: January 19, 2023).

[10] Pfeifer, R. (1988) *Artificial Intelligence models of emotion*. Available at: https://link.springer.com/chapter/10.1007/978-94-009-2792-6_12 (Accessed: January 19, 2023).

[11] Law, E. (2022) *HAIID Lectures, Durham University Computer Science.* (Accessed: January 19, 2023).

[12] Nielsen, J. (1992) *Finding usability problems through heuristic evaluation*. Available at: https://dl.acm.org/doi/10.1145/142750.142834 (Accessed: January 19, 2023).

[13] Nielsen, J. (2020) *10 Usability Heuristics for User Interface Design, Nielsen Norman Group*. Available at: https://www.nngroup.com/articles/ten-usability-heuristics/ (Accessed: January 19, 2023).

[14] ISO (2010) *Ergonomics of human-system interaction — Part 210: Human-centred design for interactive systems*. Available at: https://scc.isolutions.iso.org/obp/ui#!iso:std:iso:9241:-210:ed-2:v1:en (Accessed: January 19, 2023).

[15] Chin, D. (2000) *Empirical Evaluation of User Models and User-Adapted Systems*. Available at: https://link.springer.com/content/pdf/10.1023/A:1011127315884.pdf (Accessed: January 19, 2023).

[16] Nielsen, J. (1994) *Severity ratings for usability problems, Nielsen Norman Group*. Available at: https://www.nngroup.com/articles/how-to-rate-the-severity-of-usability-problems/ (Accessed: January 19, 2023).

[17] Nielsen, J. (2012) *How many test users in a usability study?, Nielsen Norman Group*. Available at: https://www.nngroup.com/articles/how-many-test-users/ (Accessed: January 19, 2023).

[18] Fonteyn, M., Kuipers, B. and Grobe, S. (2016) *A Description of Think Aloud Method and Protocol Analysis*. Available at: https://journals.sagepub.com/doi/10.1177/104973239300300403 (Accessed: January 19, 2023).

[19] *Guidelines for Human-AI Interaction* (2021) *Microsoft Research*. Available at: https://www.microsoft.com/en-us/research/project/guidelines-for-human-ai-interaction/ (Accessed: January 19, 2023).

[20] Kleinginna, P.R. and Kleinginna, A.M. (1981) *A categorized list of emotion definitions, with suggestions for a consensual definition*. Kluwer Academic Publishers-Plenum Publishers. Available at: https://link.springer.com/article/10.1007/BF00992553 (Accessed: January 19, 2023).

[21] Landowska, A. (2013) *Affective computing and affective learning - methods tools and prospects*. Available at: https://www.researchgate.net/publication/262574205 (Accessed: January 19, 2023).

[22] Picard, R.W. *et al.* (2004) *Affective learning — A manifesto*. Available at: https://www.media.mit.edu/publications/bttj/Paper26Pages253-269.pdf (Accessed: January 19, 2023).

[23] Tao, J. and Tan, T. (2005) *Affective Computing: A Review*. Available at: https://www.researchgate.net/publication/220270285_Affective_Computing_A_Review (Accessed: January 19, 2023).

[24] *Best Emotion Recognition Software - 2023 Reviews & Comparison* (no date). Available at: https://sourceforge.net/software/emotion-recognition/ (Accessed: January 19, 2023).

[25] *Emotimeter - Emotion Detector, Google Play*. Google. Available at: https://play.google.com/store/apps/details?id=com.reaimagine.josem.emotimeter_facialemotionrecognizer&hl=en&gl=US (Accessed: January 19, 2023).

[26] *Emotion Recognition* (2018) *Behavioral Signals*. Available at: https://behavioralsignals.com/emotion-recognition/ (Accessed: January 19, 2023).

[27] Amershi, S. *et al.* (2019) *Guidelines for Human-AI Interaction*. Available at: https://www.microsoft.com/en-us/research/uploads/prod/2019/01/Guidelines-for-Human-AI-Interaction-camera-ready.pdf (Accessed: January 19, 2023).

[28] Life, C.C.P.A.for *et al.* (2021) *Human-Centric AI: From principles to actionable and shared policies, G20 Insights*. Available at: https://www.g20-insights.org/policy_briefs/human-centric-ai-from-principles-to-actionable-and-shared-policies/ (Accessed: January 19, 2023).

[29] Nielsen, J. (2000) *Why you only need to test with 5 users, Nielsen Norman Group*. Available at: https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/ (Accessed: January 19, 2023).

[30] Hoffman, R.R., Mueller, S.T., Klein, G., and Litman, J. (2018) *Measuring Trust in the XAI Context* (Accessed: January 19, 2023).

[31] Burmester, M., Hassenzahl , M. and Koller, F. (2011) *Attrakdiff: Questionnaire*. Available at: https://www.kompetenzzentrum-usability.digital/kos/WNetz?art=File.download&id=1296&name=AttrakDiff_EN_UID.pdf (Accessed: January 19, 2023).

[32] Brooke, J. (1995) *SUS - A quick and dirty usability scale*. Available at: https://www.researchgate.net/profile/John-Brooke-

6/publication/228593520_SUS_A_quick_and_dirty_usability_scale/links/5f24381392851cd302cbaf25/SUS-A-quick-and-dirty-usability-scale.pdf (Accessed: January 19, 2023).

[33] Kujala, S. *et al.* (2011) *UX Curve: A method for evaluating long-term user experience*. Available at: https://academic.oup.com/iwc/article/23/5/473/660020 (Accessed: January 19, 2023).

[34] *Ease of Use*, *The Interaction Design Foundation*. Available at: https://www.interaction-design.org/literature/topics/ease-of-use (Accessed: January 19, 2023).

[35] *Explainability*, *IBM*. Available at: https://www.ibm.com/design/ai/ethics/explainability/ (Accessed: January 19, 2023).

[36] *Emotion detection and recognition market size*, *MarketsandMarkets*. Available at: https://www.marketsandmarkets.com/Market-Reports/emotion-detection-recognition-market-23376176.html (Accessed: January 19, 2023).

[37] Ortega, M., Rodriguez, L.-F. and Gutierrez-Garcia, J.O. (2019) *Towards emotion recognition from contextual information using machine learning*. Available at: https://link.springer.com/article/10.1007/s12652-019-01485-x (Accessed: January 19, 2023).