Annotation Practices in Affective Computing

What are these algorithms actually trained on?

Suzanne Backer[1]

Supervisors: Cynthia Liem[1]**, Andrew Demetriou**[1]

[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 24, 2023

Name of the student: Suzanne Backer
Final project course: CSE3000 Research Project
Thesis committee: Cynthia Liem, Andrew Demetriou, Frank Broz

Abstract

In the machine learning research community, significant importance is given to the optimization of techniques which are employed once a benchmark dataset is given. However, less importance is assigned to the quality of these datasets and to how these datasets are obtained. In this work, we look into annotation practices in the research area of affective computing, analysing datasets of emotion classification tasks from text, video, audio, EEG data and more. We find annotation practices of varying quality and recommend that annotation practices be improved, especially with regard to multiple annotator overlap.

1 Introduction

Applications of machine learning have an increasing potential to impact human lives today. They are able to assist in all kinds of decision-making, such as whom to grant loans to, detecting diseases or predicting consumer behaviour.

These applications are trained on or evaluated against datasets consisting of examples with a label, annotation or so-called 'ground truth' attached to them. These labels are the values the algorithm should predict if it worked accurately. As one can imagine, the quality of a machine learning application is only ever as good as the quality of the data that it was trained on. If the dataset contains inaccurately labelled examples, one can expect the predictions of the algorithm to be inaccurate as well. This phenomenon is known as 'Garbage in, garbage out'. Therefore, to obtain machine learning applications of high quality, having datasets of high quality is essential.

Datasets of high quality are datasets which are carefully annotated and accurately reflect data that can be encountered when employed in society. If datasets of low quality are used, there is the risk of data cascades. Sambasivan et al define data cascades to be 'compounding events causing negative, downstream effects from data issues, resulting in technical debt over time' [150]. After interviewing 53 AI practitioners, they found that at least 92% of high-stakes projects encountered at least one data cascade, often resulting in impacts like discarding projects, having to redo data collection or harming communities. These negative consequences are all results of the fact that work on collecting high-quality datasets isn't as appreciated as the model work [150].

1.1 Data annotations

An important factor of data quality is the quality of the annotations assigned to the data. Therefore, to prevent data cascades, it is considered of great importance to ask ourselves the question of what the quality of annotation practices in machine-learning research is. Once this question is answered, one could reason about whether this level of quality is high enough, or that greater effort should be employed to obtain and use high-quality datasets.

In an effort to gain more insight into annotation practices, Geiger et al investigated a set of machine-learning papers [37]. Specifically, they examined papers which utilized annotated data from tweets. The quality of the annotations in this set of papers was found to be diverse. This could be an indication of varying importance being assigned to the validity of annotations of a dataset.

While the work of Geiger et al is a significant contribution, the quality of annotation practices in machine learning remains a largely unexplored area of research. To obtain a clearer overview of the field, performing more reviews of different sub-domains within the machine learning community is essential.

This work contributes to our understanding of annotation practices by investigating how they are employed in the area of affective computing specifically. The research question for this work, therefore, is: 'What are current data collection and reporting practices of human annotations of machine learning research in the area of affective computing?'

1.2 Affective computing

Affective computing is defined to entail the design and implementation of systems that are able to detect, express or 'feel' an emotion [127]. Current research has mainly focused on sensing emotions through modes such as facial position, facial activity, speech, textual or electrodermal activity [30]. Often these channels have been used in unimodal analysis. However, the prediction of emotions through multimodal analysis is increasingly explored in the community of affective computing [132].

We choose to look into the annotation practices of affective computing because we consider this domain one with potentially significant societal impact. For example, affective computing could be used for monitoring and interpreting affective behavioural cues. This could be important information in the domain of security or justice, specifically for lawyers, police officers or security agents [123]. However, as one can imagine, the use of affective computing applications in these domains could have severe consequences for the people involved. Therefore it is of the utmost importance that these applications are trained on properly annotated data.

Another argument to look into affective computing, is the subjectivity of the domain. We know that humans perceive emotions differently depending on gender [47] and culture [38]. But the perception of emotion also differs from individual to individual [165]. This makes the determination of the ground truth for an emotion dataset even more challenging.

This work answers the given research question by means of a literature review of 100 works in affective computing. The goal is to focus on the most recent and highly cited papers, so as to analyse the papers with the most impact. Several questions relating to the annotation practices of the datasets these papers use will be answered. As we will see, these annotation practices are of varying quality. Both works of proper quality and of lesser quality are found. In general, we argue that annotation practices should be improved, especially with regard to multi-annotator overlap.

The remaining part of this paper is structured as follows. Chapter 2 gives an overview of the methodology employed in this study. It also gives the subquestions which were answered for each dataset encountered in the literature review.

Chapter 3 gives an overview of the answers to these questions and discusses the results of the study. Chapter 4 gives insight into how responsibly this research was conducted. Chapter 5 discusses the general findings of this study and places them in a broader context.

2 Data & Methodology

2.1 Data: a set of papers on affective computing

To analyse the annotation practices in affective computing research, our goal was to find a set of papers in this domain on which a literature review could be performed. To find these papers, inclusion and exclusion criteria for the papers were defined, after which the papers were collected and filtered.

2.1.1 Inclusion and exclusion criteria

The first criterion was that the main focus of the paper must be to either perform some kind of prediction of emotion or sentiment, or present a newly developed dataset from which this could be predicted. Papers whose main focus did not match this criterion were excluded, even when there was some overlap. An example of a paper that was excluded is one that performed facial landmark detection and eye gaze estimation but did not attempt to further process these estimations into concrete emotions [9]. Another example of a paper that was excluded was one that introduced a new natural language processing technique and evaluated this technique on multiple tasks [189]. Among these tasks was sentiment analysis, but also question answering and textual entailment. Since performing sentiment analysis was not the main focus of this study, it was not included in the set of papers for this review. To determine if this criterion was met, the abstracts of the papers were read by the author.

The second criterion was that the paper must have been published in the year 2018 or later. Since we are interested only in current annotation practices, only papers that were published in the last 5 years were taken into consideration. The third criterion concerned the type of publication. Since the aim of this study is to review annotation practices of new scientific work, publications that perform a literature review or a survey were excluded from the set of papers. The fourth criterion for the search was for the paper to be in the final stage of publication, as it was considered interesting to relate annotation practices to publication venues. So papers that were unpublished were excluded from the review. Lastly, for practical reasons, the last criterion was for the scientific work to be in English. A full list of the inclusion and exclusion criteria can be found in Table 1.

2.1.2 Data collection

To collect papers that met the given inclusion criteria, a search string was defined. To ensure that the right results would be retrieved and no relevant results would be missed by this string, several well-cited scientific works on affective computing have been read [18, 118, 132, 166]. The purpose of reading these works was to find keywords or synonyms for tasks in affective computing. Apart from reading these works, the titles of the papers that they referenced were read and scanned for relevant keywords. This led to the creation of a search string that included the keywords 'emotion* detection', 'emotion* recognition', 'emotion* analysis', 'emotion* classification', 'affect* computing', 'affect* classification', 'affect* recognition', 'affect* detection', 'affect* interpretation', 'affective state recognition', 'sentiment analysis', 'sentiment classification', 'sentiment-mining', 'sentiment mining', 'sentiment prediction', 'facial expression* detection', 'facial expression* recognition', 'facial expression* analysis', 'opinion-mining', 'opinion mining', 'opinion analysis', 'opinion classification', 'predicting emotion', 'recognizing emotion', 'recognizing affect', 'classification of affect*', 'emotional tagging', 'body gesture detection', 'body gesture recognition'. These keywords were all combined in the search string with the 'OR' operator.

The Scopus online database [1] was used to enter the search string, specifically on the date of 07-05-2023. Scopus allowed for automatic filtering for some of the inclusion and exclusion criteria, so for example papers written in languages other than English were automatically removed. Other filters that were used were the paper being published in 2018 or later and the publication stage being final. The search string was searched for in the title, abstract and keywords of the results.

2.1.3 Data filtering

The search results were processed by ordering them in descending order based on citation count. This was done because we sought a clear overview of the annotation practices in the affective computing research field. As Teplitskiy et al found, highly cited papers influence the respective research community much more than publications with lesser citations [167]. And so it followed naturally to analyse the papers with the highest citation count.

Next, the abstracts of the most cited papers were read to determine whether all inclusion criteria were met. Due to the time constraints of this study, the goal was to collect 100 papers for review.

2.2 Methodology

To analyse the annotation practices of the most cited work in affective computing, a similar approach was taken to the work from Geiger et al [37]. A list of questions was answered for each dataset included in the study. These questions were almost all taken from [37], with some additional questions added. The answers to these questions were collected by reading relevant sections of the paper itself. Often, however, relevant sections of papers that were referred to were also read in order to find the answers to the questions defined.

For each question, a set of possible answers was predefined. During the process of collecting the results, these possible answers were updated according to newly discovered possible answers. Sometimes this resulted in regrouping or redefining already collected results.

The questions that were answered are the following:

[1] https://www.scopus.com/sources.uri

	Inclusion	Exclusion
Topic	Emotion or sentiment must be estimated OR	Other
	A dataset must be introduced from which this can be estimated	
Date published	2018 or later	Before 2018
Publication type	Article, conference paper or book chapter	Review, survey
Publication stage	Final	In press
Language	English	Other

Table 1: The inclusion and exclusion criteria for scientific works in this study

1. What type of data was annotated?
2. What type of annotations were used in the paper?
3. What type of annotations were used in the original dataset?
4. Was the work an original classification task?
5. Was the dataset annotated by humans?
6. Did the paper use original human annotation?
7. Did the paper use external human annotation?
8. Who were the human annotators?
9. Was the amount of annotators specified?
10. Was the amount of annotators estimated?
11. Were formal instructions provided to the annotators?
12. Was training provided for the human annotators?
13. Was there any pre-screening for annotators from crowdwork platforms?
14. Was there multiple annotator overlap?
15. Did the paper report a metric of inter-annotator agreement?
16. Did the paper report any other metric of label quality?
17. Did the paper link to the dataset?

3 Results & Findings

3.1 Study selection

The search string as defined in section 2.1.2 resulted in 37568 matches from the Scopus database. The abstracts of the most highly cited of these papers were read until 100 papers were identified to adhere to the inclusion criteria. In the end, the abstracts of 150 papers were read. The number of papers that did not meet the inclusion criteria was 50. All papers were either excluded because their main topic was not affective computing, or they were a survey or literature review. An overview of the papers included and excluded from the study can be found in Appendix A. For most papers, other sources were also used to answer the questions as stated in Section 2.2. An overview of which sources were used for which papers can be found in Appendix B.

3.2 Amount of datasets

In this study, we will analyse the annotation practices from the perspective of the datasets instead of from that of the papers. The results for each individual dataset are made publicly available. [2] While some of the papers only used one dataset, others used multiple. The 100 papers analysed in this study used 215 datasets in total.

As all papers are in the domain of affective computing, it follows naturally that some papers use the same datasets to evaluate their work. Examples of datasets that were often used are SemEval-2014 task 4 [131], IEMOCAP [16], CK+ [98] and SEED [214]. They respectively appear 12, 11, 9 and 8 times in the total of 215 datasets.

We argue that including the same dataset multiple times is not undesirable, as the fact that the dataset is used multiple times indicates its popularity. Including it multiple times more accurately reflects the most commonly used annotation practices.

Interestingly, for 8 datasets, almost no information could be retrieved, namely 3 datasets that were used in [138] and 5 datasets that were used in [12]. Therefore, these datasets received the label 'no information' for almost all questions, except for the 'original human annotation' question. This is because it was clear that if the datasets contained human annotations, the annotations were not originally defined in the respective paper.

3.3 Type of data

Affective computing is defined as any kind of task in which emotion is predicted, regardless of the type of data used. However, one could also argue that classifying emotions from different kinds of data could be considered entirely different research fields. To have a clearer view of what type of research was analysed, the type of data that was used in a dataset was recorded. An overview can be found in Table 2.

	Count	Percentage
Text	89	41.4%
Images	47	21.86%
EEG and other physiological measures	24	11.16%
Audio and video	20	9.30%
Audio, video and text	16	7.44%
Social media content	10	4.65%
Audio	7	3.25%
Other	2	0.93%

Table 2: Type of data

Interestingly, in over 70% of cases, uni-modal datasets

[2] https://github.com/Suzanne108/Affective-Computing-Annotation-Practices

were used, showing that uni-modal affective computing is still widely researched as opposed to multimodal computing.

Next to the different types of data recorded, it is also interesting to note that the data was obtained using multiple collection methods. For datasets that contained text, often online available reviews [131] or social media posts were collected [33]. About half of the datasets which were labelled with 'text' contained reviews and the other half contained social media posts, which were often tweets.

Tweets were not categorized under 'social media content' if just the text contained in a tweet was used. In that case, they were classified as 'text'.

The datasets that were classified as 'social media content' contained more information than just the text contained in a social post. For example, they often contained information on who the social media user is connected to or to which posts the user responded. In other words, what separates the tweets in the category 'text' and the category 'social media content' is that the datasets contained a form of metadata for the 'social media content' category.

3.4 Type of annotations

Apart from what kind of data is recorded in a dataset, the types of annotations that were given to the data are interesting to analyse as well. An overview of what kind of annotations were used is given in Table 3.

The types of annotations used in the analysed papers can differ from the kind that was used in the original dataset. For example, some papers reduced a dataset that originally had 4 to 7 levels of positive and negative classification to just positive and negative. Another example is papers using a smaller amount of discrete emotions than originally provided in the dataset.

	Used annot.	Original annot.
Positive / negative	48	29
Positive / negative / neutral	36	31
Positive / negative, 4-7 levels	13	28
Discrete emotions, less than 5	15	6
Discrete emotions, 5 - 10	78	81
Discrete emotions, 10 - 15	4	8
Valence / Arousal, high / low	7	3
Valence / Arousal, range	12	24
FACS	0	2
No information	3	9

Table 3: Type of annotations. The first column indicates what annotations were used in the paper from the set of 100 analysed papers. The second column indicates what annotations the datasets originally contained. The difference indicates how many annotations were changed with regard to the original dataset. When adding the numbers in the columns, one achieves a higher number than the 215 datasets. This is because some datasets contained multiple types of annotation. FACS stands for Facial Action Coding System and refers to a system of coding facial movements in order to detect emotions.

3.5 Original classification task

Like in the original work by Geiger et al [37], an original task was defined as a machine learning algorithm or model that was novel or a novel combination of models. This does not include works which simply take an existing technique and apply it to a new dataset.

From the set of 100 papers, 92 performed an original task. From the 8 papers which were not an original task, often they performed some sentiment classification on text, such as for example in [14]. An often-used technique in these studies is to use existing lexicons, relating words to an emotion, to classify the emotion of the overall text. Since these lexicons are not original sentences or texts themselves, we do not see them as a training set and they are not labelled as such.

3.6 Labels from human annotation

The majority of datasets were labelled by human annotators (Table 4). 8 papers were labelled 'no information' as information on these datasets could not be retrieved, as explained in Section 3.2. In the following subsections, these 8 datasets will not be included in the results.

	Count	Percentage
Yes	191	88.84%
No	15	6.98%
No information	8	3.72%
Unsure	1	0.47%

Table 4: Datasets with human annotation

While for most datasets it was obvious if the labels were provided by humans, for some datasets this question was harder to answer. This is mainly because of how the different datasets were constructed.

For example, a dataset was constructed by recording participants of the study who were asked to express a certain emotion [55]. In this case, there is no explicit labelling of data. However, the human participants decided how to correctly express the emotion, so this was recorded as human annotation.

From the 215 datasets, 56 were found to record participants of a study being instructed to express an emotion. The majority, 44 datasets, also used external annotators to label the samples afterwards. Some papers provided the help of professionals to participants in expressing the emotion. For example, in [190] a psychologist explained how certain emotions should be expressed.

Other examples of edge cases include cases where the reactions of participants were recorded as they watched a video that should induce a certain emotion. From the 215 datasets, 30 were found to use this technique. A few of them also validated their choice of videos on a control group, as done in the construction of the DEAP dataset [73].

The papers that used videos to elicit emotions, mostly also asked the participants to self-assess their emotions afterwards. This was done to verify that the right emotions were elicited. 4 of them also had the recordings checked by expert annotators afterwards. But for the datasets that contained EEG data, this was harder to do.

All of the cases where emotion was elicited in some way were recorded as human annotation, even if only self-assessment was provided as the label. However, one could

argue that the quality of the labels increased as more verification steps are taken.

3.7 Original and external human annotation

Original human annotation was defined as the authors of a paper obtaining new labels for examples of a dataset. From the 100 analysed papers, 27 datasets were used which were at least partly constructed by original human annotation (Table 5). Some datasets contained both original and external labels.

	Yes	No
Original annotation	27	169
External annotation	165	26

Table 5: Used original or external human annotation

For 8 datasets for which no information could be retrieved, it was at least clear that no original human annotations were used. That is why the total amount of papers that did use original annotation and that did not is higher than the number of papers that were annotated by humans, as shown in Table 5.

3.8 Human annotation source

The most common source for human annotations was human experts (Table 6). The label 'experts' was used whenever the paper mentioned that the labellers had more knowledge than the average public. For example, some used linguists [129] or psychologists [191]. Some of the datasets that were labelled as 'experts' were both labelled by students and an expert. In contrast, the category 'students' contains datasets that were labelled only by students.

	Count	Percentage
Experts	52	27.23%
No information	30	15.71%
Paper's authors	29	15.18%
Self-assessment	23	12.04%
Students	18	9.42%
Non-experts	17	8.90%
Amazon Mechanical Turk	14	7.33%
Other crowdwork	6	3.14%
Other	2	1.05%

Table 6: Human annotation source

Defining the source of human annotations was challenging for studies that used videos or other tools to elicit emotions from participants. One could argue that the authors of the paper are the annotators, as they choose which video belonged to which emotional state. However, if self-assessment was performed after viewing such a video, the participants of the study could also be defined as the source. A separate category 'self-assesment' was created for this purpose.

3.9 Number of annotators specified or estimated

The following two questions concerned whether the number of annotators was specified and estimated. Quite some datasets specified the number of annotators, (Table 7). However, remarkably, none of them made an estimation of the required amount of annotators. It could of course still be the case that the creators of the dataset thought about this question, and made an attempt to recruit the number of annotators accordingly. However, none of this was reported.

Some datasets are labeled as 'no information' instead of 'no' here, because the original paper on the dataset could not be tracked. However, they didn't belong to the 8 as mentioned before in Section 3.2, because the original paper using them provided at least some further details on the dataset.

	Yes	No	N/A	No information
Annotators specified	136	50	2	3
Annotators estimated	0	185	2	4

Table 7: Number of annotators specified and estimated

3.10 Instructions and training for human annotators

For the instructions question, a distinction was made between annotators who have been given instructions with lengthy definitions or examples, and annotators that were just given a question. This was done because in the second case, the authors of the paper at least still provided some information on the given instructions. The majority of papers did not provide such information, (Table 8).

	Count	Percentage
Instr. with definitions or examples	69	36.12%
No instr. beyond question text	18	9.42%
No information	100	52.35%
N/A	4	2.09%

Table 8: Were instructions provided to annotators?

Most papers also did not provide any information on training for annotators, (Table 9). As opposed to instructions, training was defined as an interactive process in which annotators could receive feedback about the quality of their annotations. About 5% of datasets explicitly reported that no training was provided for the annotators.

Important to note is that for some datasets, the level of professionality of the annotators is quite high. For example, papers using psychologists as annotators [191]. Since psychologists are educated on emotions, one could argue that training is less relevant for these annotators.

	Count	Percentage
No information	132	69.11%
Some training details	45	23.56%
No	10	5.24%
N/A	4	2.09%

Table 9: Training for human annotators

3.11 Prescreening for crowdworking platforms

For the 20 datasets that were annotated with the use of crowdworking platforms, 9 of them performed some kind of screen-

ing for the annotators, (Table 10).

Authors creating a dataset using a crowdworking platform sometimes include some samples that are accurately annotated by the authors themselves. Crowdworkers are then randomly assigned to rate these as well. Sometimes, the labels created by these crowdworkers are only kept in the final dataset if their approval rate is above some certain threshold. This reflects the 'approval rate' category in Table 10.

	Count	Percentage
No information	11	55.00%
Approval rate	5	25.00%
Location qualification	3	15.00%
Projec-specific prescreening	1	5%

Table 10: Prescreening for crowdwork platforms

3.12 Multiple annotator overlap

Multiple annotator overlap was defined as multiple annotators annotating the same examples of a dataset. This was done at least partially for about 65% of the datasets, (Table 11).

	Count	Percentage
Yes, for all items	98	51.31%
Yes, for some items	25	13.09%
No	22	11.52%
No information	44	23.04%
N/A	2	1.05%

Table 11: Multiple annotator overlap

For the papers that used self-assessment or a recording of actors, multiple annotator overlap might be harder to achieve. However, since external validators of the emotions could also have been used, these papers still received the label 'no'.

One notable example of a dataset that received a 'yes' on this question was a paper that used an already annotated dataset, which did not have multiple annotator overlap, but the authors of the paper validated the labels themselves [197].

3.13 Reporting of inter-annotator agreement or another metric of label quality

Almost 60% of the datasets which performed multiple annotator overlap, reported the inter-annotator agreement, (Table 12). It is worth noting, however, that different levels of inter-annotator agreement were encountered while labelling the datasets.

	Count	Percentage
Yes	73	59.35%
No	50	40.65%

Table 12: Reported inter-annotator agreement

Often, the paper's authors also did not take any additional measures when this agreement was on the lower side. For example, the reported inter-annotator agreement for AffectNet [114] was 60.7%. In addition to that, only part of the dataset was labelled by two annotators, the majority was labelled by one person. And so, one could argue that the labels of this dataset are very dependent on the subjective interpretation of this one annotator.

Next to inter-annotator agreement, papers which used human annotation were screened for reporting some other kind of metric of label quality. The results of this analysis can be found in Table 13.

	Count	Percentage
Yes	36	19.37%
No	152	79.58%
No information	2	1.05%

Table 13: Reported some other metric of label quality

3.14 Link to the dataset provided

The last question concerned the paper providing a link to the dataset, such that it is accessible for people who might want to use the dataset. An overview of the results can be found in Table 14.

	Count	Percentage
No	88	40.93%
Yes, open source	51	23.72%
Yes, but link was broken	41	19.07%
Yes, but request access	34	15.81%
Unsure	1	0.47%

Table 14: Provided link to dataset

Often, the original 100 papers referred to other papers for more information on the datasets that were used. If the links to the datasets were provided in the linked papers, the answer to this question was recorded as 'yes' as well.

However, it also often was the case that a paper provided a link to a website which was no longer available. As seen from Table 14, this happened in almost 20% of the cases.

For papers that did provide a link to the dataset, a distinction between two categories was made. On the one hand, there were datasets that were directly available for download from the web. Other datasets could be obtained by for example filling in a request form or sending an email to the academics who created the datasets. For the second case, it often was clear that only researchers affiliated with a university would receive the dataset upon request.

4 Responsible Research

It is of utmost importance that all research is conducted ethically and responsibly. On top of that, research should always be reproducible. We will reflect on these matters related to this study in this chapter.

4.1 Ethical concerns

Firstly, one should not forget that, as in many scientific works, there is the risk of confirmation bias. If the researchers of a study assume to find a certain result, they could see it, even if it is not necessarily there. Because we originally assumed

that annotation practices would not be up to standards, there is the risk that we do not judge the findings objectively.

Secondly, there is the risk of selection bias. Because of the limited resources for this study, no more than 100 papers could be analysed. This could potentialy be too small of a sample size to conclude something about the entire field of affective computing. Still, considering that these are the most highly cited papers from the last 5 years, we are confident that this is a representative sample.

4.2 Reproducibility

The PRISMA guidelines [3] for literature review have been followed where applicable to ensure that the work presented is reproducible. However, some steps in the process are inherently subjective and might therefore not be entirely reproducible. For example, the determination of whether a paper should be included or excluded from the review, is, although the inclusion and exclusion criteria have been defined and followed, still inherently a subjective judgement. Therefore, other researchers performing this study might decide to include some studies which were excluded in this work or the other way around.

5 Discussion

5.1 Varying Quality

Based on our experience of analysing 100 papers and 215 datasets on their annotation practices, we can say that in general, the quality of the annotation practices varies greatly from dataset to dataset and from paper to paper. We found datasets that were constructed and annotated with quite some care and attention. This for example holds for the RECOLA dataset [142], which consists of audio and video from 46 French-speaking participants and was provided with annotations from 6 French-speaking annotators. The authors report that the annotators were given a document explaining the procedure of the annotation process and some explanation of emotional cues. Besides that, the paper reported that before starting the annotation process, the annotators were given examples from another database, as a form of practice. The authors also report a fairly good inter-annotator agreement. The RECOLA dataset was only used one time in the total of 215 datasets.

On the other hand, we also found quite some papers that took very little care of their annotation practices or even for which no information could be found. For example, the SFEW dataset [32], which only reported that the data was annotated by two independent labellers. However, since no inter-annotator agreement was reported, it is hard to tell if this implies multiple annotator overlap. The SFEW dataset is used four times in the total of 215 datasets.

Seeing such differences in the quality of annotation practices in the 100 most cited papers on affective computing is concerning. Especially because high-quality annotations are specifically important in affective computing, since this is a domain in which the resulting predictions could be of high societal impact.

[3] http://www.prisma-statement.org/

5.2 Multiple Annotator Overlap

Even for papers that put quite some effort into their annotation practices, there are some parts that can be improved. For example, the finding that no papers explicitly reported an estimation of the number of annotators needed (Table 7, is considered concerning. Estimating the necessary amount of annotators is important because a high enough number of annotators is needed to perform a good practice of multiple annotator overlap.

Since even humans can perceive emotions differently from each other [38, 47, 165], we consider multiple annotator overlap as one of the most important measures one could take to construct an emotion dataset with high-quality annotations. Taking this into account, we consider around 50% of datasets using multiple-annotator overlap for the whole dataset (table 11) to be a quite low amount. When we find that in turn, only 60% of those report the inter-annotator agreement, this raises even more concerns.

When using multiple annotator overlap, it is also important to recruit a large enough number of annotators to label the same items. One could imagine that a dataset that was labelled by 20 people is more reliable than one that is labelled by only 2. A limitation of this study is that this number was not recorded. However, we have come across various amounts of annotators during the collection of the results.

5.3 Other findings

Worth mentioning is that authors quite often do not report parts of the information that we are interested in. For example, for almost 15% of datasets, it is not clear who the annotators were (Table 6), and no information was provided on instructions and training details for respectively around 50% and 70% of all datasets (Table 8 and 9). For training, we can imagine that no information provided, also means no training provided. However, concerning instructions, we suspect that most authors at least provided some kind of question to the annotators. Yet, this was not reported. This is an indicator of the lack of attention that is given to the reporting of annotation practices.

Another remarkable finding is that for almost 60% of datasets, there was no available link to the dataset provided by the authors of the papers. This raises concerns about the reproducibility of studies in affective computing.

Furthermore, we find that the type of annotations which are given to datasets, but also used in practice by papers, differ a lot (Table 3). We would argue that the field of affective computing could use some unity in the annotations which are given to examples, as this gives more clarity as to what a well-performing model should exactly be able to predict.

5.4 Summary

All in all, we would argue that most of the annotation practices in affective computing are not necessarily very bad, however, considering the importance of high-quality annotations, especially in this domain, we would argue they could and should be more up to standards. We would argue that the most important point of improvement is the use of multiple annotator overlap.

Furthermore, it is remarkable that we see both examples of papers with annotation practices which would be considered up to standards, and also examples of papers which almost provide no information at all. This could be an indication of the varying importance assigned to the quality of annotations. As was already established by Sambasivan et al [150], we notice that little attention is given to collecting high-quality datasets compared to the attention which is given to building high-performing machine learning models.

We only considered publications in the domain of affective computing, and are therefore hesitant to generalise our findings to that of the whole field of machine learning. However, it is important to note that these findings are similar to those found by Geiger et al [37]. And so, the two studies together build a stronger case for machine learning datasets having varying annotation qualities.

5.5 Limitations

A limitation of this work is that the annotations for the papers and datasets were only collected by one person, the author of this paper. It is realistic to consider that if they were annotated by multiple people, with some kind of discussion on cases where there was disagreement, that datasets would have been annotated differently. However, we would argue that the general trend that can be seen from the results would remain the same.

Another limitation of this work one should consider is that the analysed datasets were constructed in varying manners. This makes comparing their annotation practices a challenging task.

6 Conclusions and Future Work

This study analysed the annotation practices of the datasets used in the 100 most cited works in affective computing. We found that the annotation practices in affective computing are of varying quality. In general, we recommend that they should be improved, mainly by making more use of multiple annotator overlap.

In general, we recommend the field of affective computing to create standards or guidelines for how data in affective computing should be collected and annotated. Having concrete standards will make it easier for researchers to adhere to these standards.

This work contributes to the understanding of annotation practices in machine learning research in general. However, to understand the entire field, more research should be conducted in different domains.

A Papers excluded and included in the study

Table 15: Overview of papers included and excluded from the study

	Publication
Excluded	
Affective computing not main topic	[125] [189] [9] [10] [44] [93] [1] [141] [6] [202] [140] [75] [28] [154] [210] [77] [80] [169] [67] [194] [48] [96] [182] [92] [49] [61] [178] [128] [212] [115] [176] [7] [59] [40]
Survey or review	[126] [72] [60] [184] [84] [70] [201] [192] [31] [34] [13] [78] [24] [54] [164] [135]
Included	[85] [114] [211] [100] [112] [91] [159] [197] [69] [215] [193] [213] [183] [83] [152] [12] [186] [19] [174] [68] [89] [20] [26] [27] [35] [103] [180] [172] [199] [173] [56] [203] [206] [161] [162] [151] [134] [137] [52] [181] [58] [5] [87] [139] [101] [17] [50] [138] [105] [86] [163] [204] [21] [23] [36] [102] [187] [55] [157] [14] [62] [148] [121] [175] [79] [94] [63] [65] [22] [188] [200] [116] [90] [216] [155] [109] [4] [205] [95] [51] [2] [53] [46] [82] [71] [64] [81] [39] [170] [108] [110] [41] [66] [168] [25] [145] [76] [74] [117]

B Other sources used for collecting results

Table 16: Other sources used for collecting results per paper part 1

Paper	Other sources used
[211]	[15, 16]
[100]	[130, 146]
[91]	[32, 83, 98, 114, 124, 209]
[159]	[69, 214]
[197]	[98]
[215]	[73, 214]
[183]	[129–131]
[12]	[42, 171]
[186]	[98, 124, 190, 191, 207–209]
[19]	[119, 120, 143, 147, 156]
[174]	[11, 32, 83, 114]
[68]	[42, 144, 147]
[89]	[33, 131]
[20]	[133, 156]
[26]	[15, 16]
[35]	[33, 131]
[103]	[16, 107]
[172]	[179]
[199]	[33, 129–131]
[173]	[11, 83, 114]

Table 17: Other sources used for collecting results per paper part 2

Paper	Other sources used
[56]	[104]
[203]	[15, 104, 177, 198]
[206]	[98, 214]
[162]	[131, 146]
[52]	[16]
[181]	[129, 131]
[58]	[131]
[5]	[129]
[87]	[88, 185]
[139]	[57, 122, 156]
[101]	[130, 146]
[17]	[32, 98, 124, 209]
[50]	[73]
[117]	[15, 16, 97]
[86]	[73, 214]
[163]	[33, 129, 131]
[204]	[104, 177, 198]
[36]	[33, 131]
[102]	[16, 195]
[55]	[104]
[121]	[33]
[62]	[15, 16, 97]
[175]	[33, 131]
[79]	[73, 214]
[63]	[98, 99]
[22]	[73]
[188]	[73]
[200]	[32, 45, 191]
[116]	[16, 97]
[90]	[33, 129–131]
[109]	[8, 43, 98, 99]
[4]	[29, 111]
[95]	[106]
[51]	[16, 107]
[53]	[129–131]
[46]	[73, 214]
[82]	[73, 158, 214]
[71]	[98, 99]
[64]	[99, 124]
[81]	[214]
[39]	[11, 43, 114]
[170]	[142]
[108]	[15, 16]
[41]	[16, 153]
[168]	[3, 149, 196]
[25]	[98, 99]
[76]	[112, 113, 136, 160]

References

[1] Ala Abd-Alrazaq, Dari Alhuwail, Mowafa Househ, Mounir Hai, and Zubair Shah. Top concerns of tweeters during the COVID-19 pandemic: A surveillance study. *Journal of Medical Internet Research*, 22(4), 2020. Cited by: 442; All Open Access, Gold Open Access, Green Open Access.

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

[3] Nawaf A. Abdulla, Nizar A. Ahmed, Mohammed A. Shehab, and Mahmoud Al-Ayyoub. Arabic sentiment analysis: Lexicon-based and corpus-based. In *2013 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)*, pages 1–6, 2013.

[4] Md Shad Akhtar, Asif Ekbal, and Erik Cambria. How intense are you? Predicting intensities of emotions and sentiments using stacked ensemble [application notes]. *IEEE Computational Intelligence Magazine*, 15(1):64–75, 2020.

[5] Mohammad Al-Smadi, Omar Qawasmeh, Mahmoud Al-Ayyoub, Yaser Jararweh, and Brij Gupta. Deep recurrent neural network vs. support vector machine for aspect-based sentiment analysis of Arabic hotels' reviews. *Journal of Computational Science*, 27:386–393, 2018.

[6] Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2890–2896, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[7] Halima Amjad, David L. Roth, Orla C. Sheehan, Constantine G. Lyketsos, Jennifer L. Wolff, and Quincy M. Samus. Underdiagnosis of dementia: an observational study of patterns in diagnosis and awareness in US older adults. *Journal of General Internal Medicine*, 33(7):1131 – 1138, 2018. Cited by: 160; All Open Access, Bronze Open Access, Green Open Access.

[8] Deepali Aneja, Alex Colburn, Gary Faigin, Linda G. Shapiro, and Barbara Mones. Modeling stylized character expressions via deep learning. In *Asian Conference on Computer Vision*, 2016.

[9] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 59–66, 2018.

[10] Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M. Martinez, and Seth D. Pollak. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, 20(1):1 – 68, 2019. Cited by: 607; All Open Access, Bronze Open Access, Green Open Access.

[11] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd-sourced label

distribution. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ICMI '16, page 279–283, New York, NY, USA, 2016. Association for Computing Machinery.

[12] Mohammad Ehsan Basiri, Shahla Nemati, Moloud Abdar, Erik Cambria, and U. Rajendra Acharya. ABCDM: An attention-based bidirectional CNN-RNN deep model for sentiment analysis. *Future Generation Computer Systems*, 115:279–294, 2021.

[13] Marouane Birjali, Mohammed Kasri, and Abderrahim Beni-Hssane. A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, 226:107134, 2021.

[14] Sakun Boon-Itt and Yukolpat Skunkan. Public perception of the COVID-19 pandemic on twitter: Sentiment analysis and topic modeling study. *JMIR Public Health and Surveillance*, 6(4), 2020. Cited by: 184; All Open Access, Gold Open Access, Green Open Access.

[15] Felix Burkhardt, Astrid Paeschke, M. Rolfes, Walter Sendlmeier, and Benjamin Weiss. A database of German emotional speech. volume 5, pages 1517–1520, 09 2005.

[16] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower Provost, Samuel Kim, Jeannette Chang, Sungbok Lee, and Shrikanth Narayanan. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359, 12 2008.

[17] Jie Cai, Zibo Meng, Ahmed Shehab Khan, Zhiyuan Li, James O'Reilly, and Yan Tong. Island loss for learning discriminative features in facial expression recognition. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 302–309, 2018.

[18] Erik Cambria. Affective Computing and Sentiment Analysis. *IEEE Intelligent Systems*, 31(2):102–107, 2016.

[19] Erik Cambria, Yang Li, Frank Z. Xing, Soujanya Poria, and Kenneth Kwok. SenticNet 6: Ensemble application of symbolic and subsymbolic AI for sentiment analysis. page 105 – 114, 2020. Cited by: 303.

[20] Erik Cambria, Soujanya Poria, Devamanyu Hazarika, and Kenneth Kwok. SenticNet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings. page 1795 – 1802, 2018. Cited by: 286.

[21] Koyel Chakraborty, Surbhi Bhatia, Siddhartha Bhattacharyya, Jan Platos, Rajib Bag, and Aboul Ella Hassanien. Sentiment analysis of COVID-19 tweets by deep learning classifiers—a study to show how popularity is affecting accuracy in social media. *Applied Soft Computing*, 97:106754, 2020.

[22] Hao Chao, Liang Dong, Yongli Liu, and Baoyun Lu. Emotion recognition from multiband eeg signals using capsnet. *Sensors (Switzerland)*, 19(9), 2019. Cited by: 174; All Open Access, Gold Open Access, Green Open Access.

[23] Ankush Chatterjee, Umang Gupta, Manoj Kumar Chinnakotla, Radhakrishnan Srikanth, Michel Galley, and Puneet Agrawal. Understanding emotions in text using deep learning and big data. *Computers in Human Behavior*, 93:309–317, 2019.

[24] Iti Chaturvedi, Erik Cambria, Roy E. Welsch, and Francisco Herrera. Distinguishing between facts and opinions for sentiment analysis: Survey and challenges. *Information Fusion*, 44:65–77, 2018.

[25] Luefeng Chen, Mengtian Zhou, Wanjuan Su, Min Wu, Jinhua She, and Kaoru Hirota. Softmax regression based deep sparse autoencoder network for facial emotion recognition in human-robot interaction. *Information Sciences*, 428:49–61, 2018.

[26] Mingyi Chen, Xuanji He, Jing Yang, and Han Zhang. 3-D convolutional recurrent neural networks with attention model for speech emotion recognition. *IEEE Signal Processing Letters*, 25(10):1440–1444, 2018.

[27] Mingming Cheng and Xin Jin. What do Airbnb users care about? an analysis of online review comments. *International Journal of Hospitality Management*, 76:58–70, 2019.

[28] Leon Ciechanowski, Aleksandra Przegalinska, Mikolaj Magnuski, and Peter Gloor. In the shades of the uncanny valley: An experimental study of human–chatbot interaction. *Future Generation Computer Systems*, 92:539–548, 2019.

[29] Keith Cortis, André Freitas, Tobias Daudert, Manuela Huerlimann, Manel Zarrouk, Siegfried Handschuh, and Brian Davis. SemEval-2017 task 5: Fine-grained sentiment analysis on financial microblogs and news. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 519–535, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[30] B. Daily Shaundra. Affective Computing: Historical Foundations, Current Applications, and Future Trends. *Emotions and Affect in Human Factors and Human-Computer Interaction*, page 213, 2017.

[31] Nhan Cach Dang, María N. Moreno-García, and Fernando De la Prieta. Sentiment analysis based on deep learning: A comparative study. *Electronics (Switzerland)*, 9(3), 2020. Cited by: 201; All Open Access, Gold Open Access, Green Open Access.

[32] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 2106–2112, 2011.

[33] Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. Adaptive recursive neural network for target-dependent Twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of*

the Association for Computational Linguistics (Volume 2: Short Papers), pages 49–54, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[34] Maria Egger, Matthias Ley, and Sten Hanke. Emotion recognition from physiological signal analysis: A review. *Electronic Notes in Theoretical Computer Science*, 343:35 – 55, 2019. Cited by: 179; All Open Access, Gold Open Access.

[35] Feifan Fan, Yansong Feng, and Dongyan Zhao. Multi-grained attention network for aspect-level sentiment classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3433–3442, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[36] Zhengjie Gao, Ao Feng, Xinyu Song, and Xi Wu. Target-dependent sentiment classification with BERT. *IEEE Access*, 7:154290–154299, 2019.

[37] R. S. Geiger, K. Yu, Y. L. Yang, M. Dai, J. Qiu, R. Tang, and J. Huang. Garbage in, garbage out? Do machine learning application papers in social computing report where human-labeled training data comes from? *Fat* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 325–336, 2020. Bq8fj Times Cited:28 Cited References Count:68.

[38] Maria Gendron, Debi Roberson, Jacoba Marietta van der Vyver, and Lisa Feldman Barrett. Perceptions of emotion from facial expressions are not culturally universal: Evidence from a remote culture. *Emotion*, 14(2):251 – 262, 2014. Cited by: 216; All Open Access, Green Open Access.

[39] Mariana-Iuliana Georgescu, Radu Tudor Ionescu, and Marius Popescu. Local learning with deep and handcrafted features for facial expression recognition. *IEEE Access*, 7:64827–64836, 2019.

[40] Bissan Ghaddar and Joe Naoum-Sawaya. High dimensional data classification and feature selection using support vector machines. *European Journal of Operational Research*, 265(3):993 – 1004, 2018. Cited by: 154.

[41] Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 154–164, Hong Kong, China, November 2019. Association for Computational Linguistics.

[42] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *Processing*, 150, 01 2009.

[43] Ian J. Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, Yingbo Zhou, Chetan Ramaiah, Fangxiang Feng, Ruifan Li, Xiaojie Wang, Dimitris Athanasakis, John Shawe-Taylor, Maxim Milakov, John Park, Radu Ionescu, Marius Popescu, Cristian Grozea, James Bergstra, Jingjing Xie, Lukasz Romaszko, Bing Xu, Zhang Chuang, and Yoshua Bengio. Challenges in representation learning: A report on three machine learning contests. *Neural Networks*, 64:59–63, 2015. Special Issue on "Deep Learning of Representations".

[44] Allison J. Greaney, Andrea N. Loes, Katharine H.D. Crawford, Tyler N. Starr, Keara D. Malone, Helen Y. Chu, and Jesse D. Bloom. Comprehensive mapping of mutations in the SARS-CoV-2 receptor-binding domain that affect recognition by polyclonal human plasma antibodies. *Cell Host Microbe*, 29(3):463–476.e6, 2021.

[45] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-PIE. *Image and Vision Computing*, 28(5):807–813, 2010. Best of Automatic Face and Gesture Recognition 2008.

[46] Vipin Gupta, Mayur Dahyabhai Chopda, and Ram Bilas Pachori. Cross-subject emotion recognition using flexible analytic wavelet transform from EEG signals. *IEEE Sensors Journal*, 19(6):2266–2274, 2019.

[47] Elizabeth Hampson, Sari M. van Anders, and Lucy I. Mullin. A female advantage in the recognition of emotional facial expressions: test of an evolutionary hypothesis. *Evolution and Human Behavior*, 27(6):401–416, 2006.

[48] Jochen Hartmann, Juliana Huppertz, Christina Schamp, and Mark Heitmann. Comparing automated text classification methods. *International Journal of Research in Marketing*, 36(1):20–38, 2019.

[49] Abdalraouf Hassan and Ausif Mahmood. Convolutional recurrent deep learning model for sentence classification. *IEEE Access*, 6:13949–13957, 2018.

[50] Mohammad Mehedi Hassan, Md. Golam Rabiul Alam, Md. Zia Uddin, Shamsul Huda, Ahmad Almogren, and Giancarlo Fortino. Human emotion recognition using deep belief network architecture. *Information Fusion*, 51:10–18, 2019.

[51] Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. ICoN: Interactive conversational memory network for multimodal emotion detection. page 2594 – 2604, 2018. Cited by: 162.

[52] Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2122–2132, New Orleans,

Louisiana, June 2018. Association for Computational Linguistics.

[53] Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. Effective attention modeling for aspect-level sentiment classification. page 1121 – 1131, 2018. Cited by: 157.

[54] Fatemeh Hemmatian and Mohammad Karim Sohrabi. A survey on classification techniques for opinion mining and sentiment analysis. *Artificial Intelligence Review*, 52(3):1495 – 1545, 2019. Cited by: 172.

[55] M. Shamim Hossain and Ghulam Muhammad. Emotion-aware connected healthcare big data towards 5G. *IEEE Internet of Things Journal*, 5(4):2399–2406, 2018.

[56] M. Shamim Hossain and Ghulam Muhammad. Emotion recognition using deep learning approach from audio–visual emotional big data. *Information Fusion*, 49:69–78, 2019.

[57] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. pages 168–177, 08 2004.

[58] Binxuan Huang, Yanglan Ou, and Kathleen M. Carley. Aspect level sentiment classification with attention-over-attention neural networks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10899 LNCS:197 – 206, 2018. Cited by: 205; All Open Access, Green Open Access.

[59] Amir Hussain and Erik Cambria. Semi-supervised learning for big social data analysis. *Neurocomputing*, 275:1662 – 1673, 2018. Cited by: 153; All Open Access, Green Open Access.

[60] Doaa Mohey El-Din Mohamed Hussein. A survey on sentiment analysis challenges. *Journal of King Saud University - Engineering Sciences*, 30(4):330–338, 2018.

[61] Md. Rafiqul Islam, Muhammad Ashad Kabir, Ashir Ahmed, Abu Raihan M. Kamal, Hua Wang, and Anwaar Ulhaq. Depression detection from social network data using machine learning techniques. *Health Information Science and Systems*, 6(1), 2018. Cited by: 177; All Open Access, Green Open Access.

[62] Dias Issa, M. Fatih Demirci, and Adnan Yazici. Speech emotion recognition with deep convolutional neural networks. *Biomedical Signal Processing and Control*, 59:101894, 2020.

[63] Deepak Kumar Jain, Pourya Shamsolmoali, and Paramjit Sehdev. Extended deep neural network for facial emotion recognition. *Pattern Recognition Letters*, 120:69–74, 2019.

[64] Neha Jain, Shishir Kumar, Amit Kumar, Pourya Shamsolmoali, and Masoumeh Zareapoor. Hybrid deep neural networks for face emotion recognition. *Pattern Recognition Letters*, 115:101–106, 2018. Multimodal Fusion for Pattern Recognition.

[65] Hamed Jelodar, Yongli Wang, Rita Orji, and Shucheng Huang. Deep sentiment classification and topic discovery on novel coronavirus or COVID-19 online discussions: NLP using LSTM recurrent neural network approach. *IEEE Journal of Biomedical and Health Informatics*, 24(10):2733–2742, 2020.

[66] Byeongki Jeong, Janghyeok Yoon, and Jae-Min Lee. Social media mining for product planning: A product opportunity mining approach based on topic modeling and sentiment analysis. *International Journal of Information Management*, 48:280–290, 2019.

[67] Britta L Jewell, Edinah Mudimu, John Stover, Debra ten Brink, Andrew N Phillips, Jennifer A Smith, Rowan Martin-Hughes, Yu Teng, Robert Glaubius, Severin Guy Mahiane, Loveleen Bansi-Matharu, Isaac Taramusi, Newton Chagoma, Michelle Morrison, Meg Doherty, Kimberly Marsh, Anna Bershteyn, Timothy B Hallett, and Sherrie L Kelly. Potential effects of disruption to HIV programmes in sub-saharan africa caused by COVID-19: results from multiple mathematical models. *The Lancet HIV*, 7(9):e629–e640, 2020.

[68] Zhao Jianqiang, Gui Xiaolin, and Zhang Xuejun. Deep convolution neural networks for twitter sentiment analysis. *IEEE Access*, 6:23253–23260, 2018.

[69] Stamos Katsigiannis and Naeem Ramzan. DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices. *IEEE Journal of Biomedical and Health Informatics*, 22(1):98–107, 2018.

[70] Ruhul Amin Khalil, Edward Jones, Mohammad Inayatullah Babar, Tariqullah Jan, Mohammad Haseeb Zafar, and Thamer Alhussain. Speech emotion recognition using deep learning techniques: A review. *IEEE Access*, 7:117327–117345, 2019.

[71] Ji-Hae Kim, Byung-Gyu Kim, Partha Pratim Roy, and Da-Mi Jeong. Efficient facial expression recognition algorithm based on hierarchical deep neural network structure. *IEEE Access*, 7:41273–41285, 2019.

[72] Byoung Chul Ko. A brief review of facial emotion recognition based on visual information. *Sensors (Switzerland)*, 18(2), 2018. Cited by: 348; All Open Access, Gold Open Access, Green Open Access.

[73] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. DEAP: A database for emotion analysis ;using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, 2012.

[74] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A. Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision*, 127(6-7):907 –

929, 2019. Cited by: 145; All Open Access, Green Open Access, Hybrid Gold Open Access.

[75] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D. Plumbley. PANNs: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 28:2880 – 2894, 2020. Cited by: 247; All Open Access, Green Open Access.

[76] Bernhard Kratzwald, Suzana Ilić, Mathias Kraus, Stefan Feuerriegel, and Helmut Prendinger. Deep learning for affective computing: Text-based emotion recognition in decision support. *Decision Support Systems*, 115:24–35, 2018.

[77] Sneha Kudugunta and Emilio Ferrara. Deep neural networks for bot detection. *Information Sciences*, 467:312 – 322, 2018. Cited by: 226; All Open Access, Bronze Open Access, Green Open Access.

[78] Ashu Kumar, Amandeep Kaur, and Munish Kumar. Face detection techniques: a review. *Artificial Intelligence Review*, 52(2):927 – 948, 2019. Cited by: 169.

[79] Zirui Lan, Olga Sourina, Lipo Wang, Reinhold Scherer, and Gernot R. Müller-Putz. Domain adaptation techniques for EEG-based emotion recognition: A comparative study on two public datasets. *IEEE Transactions on Cognitive and Developmental Systems*, 11(1):85–94, 2019.

[80] Jinmook Lee, Changhyeon Kim, Sanghoon Kang, Dongjoo Shin, Sangyeob Kim, and Hoi-Jun Yoo. UNPU: A 50.6tops/w unified deep neural network accelerator with 1b-to-16b fully-variable weight bit-precision. volume 61, page 218 – 220, 2018. Cited by: 217.

[81] Jinpeng Li, Zhaoxiang Zhang, and Huiguang He. Hierarchical convolutional neural networks for EEG-based emotion recognition. *Cognitive Computation*, 10(2):368 – 380, 2018. Cited by: 159.

[82] Peiyang Li, Huan Liu, Yajing Si, Cunbo Li, Fali Li, Xuyang Zhu, Xiaoye Huang, Ying Zeng, Dezhong Yao, Yangsong Zhang, and Peng Xu. EEG based emotion recognition by combining functional connectivity network and local activations. *IEEE Transactions on Biomedical Engineering*, 66(10):2869–2881, 2019.

[83] Shan Li and Weihong Deng. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Transactions on Image Processing*, 28(1):356–370, 2019.

[84] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, 13(3):1195 – 1215, 2022. Cited by: 241; All Open Access, Green Open Access.

[85] Sijia Li, Yilin Wang, Jia Xue, Nan Zhao, and Tingshao Zhu. The impact of covid-19 epidemic declaration on psychological consequences: A study on active weibo users. *International Journal of Environmental Research and Public Health*, 17(6), 2020.

[86] Xiang Li, Dawei Song, Peng Zhang, Yazhou Zhang, Yuexian Hou, and Bin Hu. Exploring EEG features in cross-subject emotion recognition. *Frontiers in Neuroscience*, 12(MAR), 2018. Cited by: 192; All Open Access, Gold Open Access, Green Open Access.

[87] Xiaobai Li, Xiaopeng Hong, Antti Moilanen, Xiaohua Huang, Tomas Pfister, Guoying Zhao, and Matti Pietikainen. Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods. *IEEE Transactions on Affective Computing*, 9(4):563 – 577, 2018. Cited by: 199; All Open Access, Green Open Access.

[88] Xiaobai Li, Tomas Pfister, Xiaohua Huang, Guoying Zhao, and Matti Pietikäinen. A spontaneous micro-expression database: Inducement, collection and baseline. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–6, 2013.

[89] Xin Li, Lidong Bing, Wai Lam, and Bei Shi. Transformation networks for target-oriented sentiment classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 946–956, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[90] Xin Li, Lidong Bing, Piji Li, and Wai Lam. A unified model for opinion target extraction and target sentiment prediction. page 6714 – 6721, 2019. Cited by: 170.

[91] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Occlusion aware facial expression recognition using CNN with attention mechanism. *IEEE Transactions on Image Processing*, 28(5):2439–2450, 2019.

[92] Andy T. Liu, Shu-Wen Yang, Po-Han Chi, Po-Chun Hsu, and Hung-Yi Lee. Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders. volume 2020-May, page 6419 – 6423, 2020. Cited by: 182; All Open Access, Green Open Access.

[93] Gang Liu and Jiabao Guo. Bidirectional LSTM with attention mechanism and convolutional layer for text classification. *Neurocomputing*, 337:325–338, 2019.

[94] Yong-Jin Liu, Minjing Yu, Guozhen Zhao, Jinjing Song, Yan Ge, and Yuanchun Shi. Real-time movie-induced discrete emotion recognition from EEG signals. *IEEE Transactions on Affective Computing*, 9(4):550–562, 2018.

[95] Zhen-Tao Liu, Min Wu, Wei-Hua Cao, Jun-Wei Mao, Jian-Ping Xu, and Guan-Zheng Tan. Speech emotion recognition based on feature selection and extreme learning machine decision tree. *Neurocomputing*, 273:271–280, 2018.

[96] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. volume 1, page

2247 – 2256, 2018. Cited by: 194; All Open Access, Green Open Access, Hybrid Gold Open Access.

[97] Steven Livingstone and Frank Russo. The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLOS ONE*, 13:e0196391, 05 2018.

[98] Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pages 94–101, 2010.

[99] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba. Coding facial expressions with Gabor wavelets. In *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, pages 200–205, 1998.

[100] Yukun Ma, Haiyun Peng, and Erik Cambria. Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive LSTM. page 5876 – 5883, 2018. Cited by: 427.

[101] Yukun Ma, Haiyun Peng, Tahir Khan, Erik Cambria, and Amir Hussain. Sentic LSTM: a hybrid network for targeted aspect-based sentiment analysis. *Cognitive Computation*, 10(4):639 – 650, 2018. Cited by: 200.

[102] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, and S. Poria. Multimodal sentiment analysis using hierarchical fusion with context modeling. *Knowledge-Based Systems*, 161:124–133, 2018.

[103] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. DialogueRNN: An attentive RNN for emotion detection in conversations. page 6818 – 6825, 2019. Cited by: 272.

[104] O. Martin, I. Kotsia, B. Macq, and I. Pitas. The eNTERFACE' 05 audio-visual emotion database. In *22nd International Conference on Data Engineering Workshops (ICDEW'06)*, pages 8–8, 2006.

[105] Javier Marín-Morales, Juan Luis Higuera-Trujillo, Alberto Greco, Jaime Guixeres, Carmen Llinares, Enzo Pasquale Scilingo, Mariano Alcañiz, and Gaetano Valenza. Affective computing in virtual reality: emotion recognition from brain and heartbeat dynamics using wearable sensors. *Scientific Reports*, 8(1), 2018. Cited by: 192; All Open Access, Gold Open Access, Green Open Access.

[106] Jindřich Matoušek, Josef Psutka, and Jiri Kruta. Design of speech corpus for text-to-speech synthesis. pages 2047–2050, 09 2001.

[107] Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3(1):5–17, 2012.

[108] Hao Meng, Tianhao Yan, Fei Yuan, and Hongwei Wei. Speech emotion recognition from 3D log-mel spectrograms with deep learning network. *IEEE Access*, 7:125868–125881, 2019.

[109] Shervin Minaee, Mehdi Minaei, and Amirali Abdolrashidi. Deep-emotion: Facial expression recognition using attentional convolutional network. *Sensors*, 21(9), 2021. Cited by: 168; All Open Access, Gold Open Access, Green Open Access.

[110] Juan Abdon Miranda-Correa, Mojtaba Khomami Abadi, Nicu Sebe, and Ioannis Patras. AMIGOS: A dataset for affect, personality and mood research on individuals and groups. *IEEE Transactions on Affective Computing*, 12(2):479–493, 2021.

[111] Saif Mohammad and Felipe Bravo-Marquez. WASSA-2017 shared task on emotion intensity. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 34–49, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[112] Saif M. Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. SemEval-2018 task 1: Affect in tweets. page 1 – 17, 2018. Cited by: 430.

[113] Saif M. Mohammad, Xiaodan Zhu, Svetlana Kiritchenko, and Joel Martin. Sentiment, emotion, purpose, and style in electoral tweets. *Information Processing Management*, 51(4):480–499, 2015.

[114] Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. AffectNet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2019.

[115] Igor Mordatch and Pieter Abbeel. Emergence of grounded compositional language in multi-agent populations. page 1495 – 1502, 2018. Cited by: 163.

[116] Mustaqeem and Soonil Kwon. A CNN-assisted enhanced audio signal processing for speech emotion recognition. *Sensors (Switzerland)*, 20(1), 2020. Cited by: 171; All Open Access, Gold Open Access, Green Open Access.

[117] Mustaqeem, Muhammad Sajjad, and Soonil Kwon. Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM. *IEEE Access*, 8:79861 – 79875, 2020. Cited by: 154; All Open Access, Gold Open Access.

[118] Mika V. Mäntylä, Daniel Graziotin, and Miikka Kuutila. The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. *Computer Science Review*, 27:16–32, 2018.

[119] Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. SemEval-2016 Task

4: Sentiment Analysis in Twitter. *arXiv e-prints*, page arXiv:1912.01973, December 2019.

[120] Preslav Nakov, Sara Rosenthal, Zornitsa Kozareva, Veselin Stoyanov, Alan Ritter, and Theresa Wilson. SemEval-2013 task 2: Sentiment analysis in Twitter. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 312–320, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics.

[121] Aytuğ Onan. Sentiment analysis on product reviews based on weighted word embeddings and deep neural networks. *Concurrency and Computation: Practice and Experience*, 33(23), 2021. Cited by: 182.

[122] Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, page 115–124, USA, 2005. Association for Computational Linguistics.

[123] M. Pantic and L.J.M. Rothkrantz. Toward an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE*, 91(9):1370–1390, 2003.

[124] M. Pantic, M. Valstar, R. Rademaker, and L. Maat. Web-based database for facial expression analysis. In *2005 IEEE International Conference on Multimedia and Expo*, pages 5 pp.–, 2005.

[125] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[126] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. volume 1, page 2227 – 2237, 2018. Cited by: 5227.

[127] Rosalind W. Picard. *Affective Computing*. The MIT Press, 07 2000.

[128] Amy E Pinkham, Philip D Harvey, and David L Penn. Social Cognition Psychometric Evaluation: Results of the Final Validation Study. *Schizophrenia Bulletin*, 44(4):737–748, 08 2017.

[129] Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California, June 2016. Association for Computational Linguistics.

[130] Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. SemEval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495, Denver, Colorado, June 2015. Association for Computational Linguistics.

[131] Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland, August 2014. Association for Computational Linguistics.

[132] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125, 2017.

[133] Soujanya Poria, Erik Cambria, Alexander Gelbukh, Federica Bisio, and Amir Hussain. Sentiment data flow analysis by means of dynamic linguistic patterns. *IEEE Computational Intelligence Magazine*, 10(4):26–36, 2015.

[134] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. page 527 – 536, 2020. Cited by: 218.

[135] Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 7:100943 – 100953, 2019. Cited by: 150; All Open Access, Gold Open Access, Green Open Access.

[136] Daniel Preoţiuc-Pietro, H. Andrew Schwartz, Gregory Park, Johannes Eichstaedt, Margaret Kern, Lyle Ungar, and Elisabeth Shulman. Modelling valence and arousal in Facebook posts. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 9–15, San Diego, California, June 2016. Association for Computational Linguistics.

[137] J. Rexiline Ragini, P.M. Rubesh Anand, and Vidhyacharan Bhaskar. Big data analytics for disaster response and recovery through sentiment analysis. *International Journal of Information Management*, 42:13–24, 2018.

[138] Guozheng Rao, Weihang Huang, Zhiyong Feng, and Qiong Cong. LSTM with sentence representations for

document-level sentiment classification. *Neurocomputing*, 308:49–57, 2018.

[139] Seyed Mahdi Rezaeinia, Rouhollah Rahmani, Ali Ghodsi, and Hadi Veisi. Sentiment analysis based on improved pre-trained word embeddings. *Expert Systems with Applications*, 117:139–147, 2019.

[140] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Semantically equivalent adversarial rules for debugging NLP models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 856–865, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[141] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online, July 2020. Association for Computational Linguistics.

[142] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8, 2013.

[143] Sara Rosenthal, Preslav Nakov, Svetlana Kiritchenko, Saif Mohammad, Alan Ritter, and Veselin Stoyanov. SemEval-2015 task 10: Sentiment analysis in Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 451–463, Denver, Colorado, June 2015. Association for Computational Linguistics.

[144] Sara Rosenthal, Alan Ritter, Preslav Nakov, and Veselin Stoyanov. SemEval-2014 task 9: Sentiment analysis in twitter. pages 73–80, 01 2014.

[145] Gonzalo A. Ruz, Pablo A. Henríquez, and Aldo Mascareño. Sentiment analysis of twitter data during critical events through Bayesian networks classifiers. *Future Generation Computer Systems*, 106:92–104, 2020.

[146] Marzieh Saeidi, Guillaume Bouchard, Maria Liakata, and Sebastian Riedel. SentiHood: Targeted aspect based sentiment analysis dataset for urban neighbourhoods. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1546–1556, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee.

[147] Hassan Saif, Miriam Fernandez, and Harith Alani. Evaluation datasets for twitter sentiment analysis. A survey and a new dataset, the STS-Gold. volume 1096, 12 2013.

[148] Kashfia Sailunaz and Reda Alhajj. Emotion and sentiment analysis from twitter text. *Journal of Computational Science*, 36:101003, 2019.

[149] Mohammed Saleh, Maria Martín-Valdivia, L. López, and José Perea-Ortega. OCA: Opinion corpus for Arabic. *JASIST*, 62:2045–2054, 10 2011.

[150] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. "Everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[151] Jim Samuel, G. G. Md. Nawaz Ali, Md. Mokhlesur Rahman, Ek Esawi, and Yana Samuel. COVID-19 public sentiment insights and machine learning for tweets classification. *Information*, 11(6), 2020.

[152] Philip Schmidt, Attila Reiss, Robert Duerichen, and Kristof Van Laerhoven. Introducing WeSAD, a multimodal dataset for wearable stress and affect detection. page 400 – 408, 2018. Cited by: 335.

[153] Björn Schuller, Michel Valstar, Roddy Cowie, and Maja Pantic. AVEC 2012 – the continuous audio/visual emotion challenge. 10 2012.

[154] Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online, November 2020. Association for Computational Linguistics.

[155] Akshit Singh, Nagesh Shukla, and Nishikant Mishra. Social media data analytics to improve supply chain management in food industries. *Transportation Research Part E: Logistics and Transportation Review*, 114:398–415, 2018.

[156] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

[157] Sahar Sohangir, Dingding Wang, Anna Pomeranets, and Taghi M. Khoshgoftaar. Big Data: Deep learning for financial sentiment analysis. *Journal of Big Data*, 5(1), 2018. Cited by: 184; All Open Access, Gold Open Access.

[158] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, 3(1):42–55, 2012.

[159] Tengfei Song, Wenming Zheng, Peng Song, and Zhen Cui. EEG emotion recognition using dynamical graph convolutional neural networks. *IEEE Transactions on Affective Computing*, 11(3):532–541, 2020.

[160] Carlo Strapparava and Rada Mihalcea. SemEval-2007 task 14: Affective text. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, page 70–74, USA, 2007. Association for Computational Linguistics.

[161] Ramanathan Subramanian, Julia Wache, Mojtaba Khomami Abadi, Radu L. Vieriu, Stefan Winkler, and Nicu Sebe. ASCERTAIN: Emotion and personality recognition using commercial sensors. *IEEE Transactions on Affective Computing*, 9(2):147–160, 2018.

[162] Chi Sun, Luyao Huang, and Xipeng Qiu. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. volume 1, page 380 – 385, 2019. Cited by: 232.

[163] Kai Sun, Richong Zhang, Samuel Mensah, Yongyi Mao, and Xudong Liu. Aspect-level sentiment analysis via convolution over dependency tree. page 5679 – 5688, 2019. Cited by: 192.

[164] Monorama Swain, Aurobinda Routray, and P. Kabisatpathy. Databases, features and classifiers for speech emotion recognition: a review. *International Journal of Speech Technology*, 21(1):93 – 120, 2018. Cited by: 167.

[165] R.H. Swain, A.J. O'Hare, and K. Brandley. Individual differences in social intelligence and perception of emotion expression of masked and unmasked faces. *Cogn Research*, 7(54), 2022.

[166] Jianhua Tao and Tieniu Tan. *Affective Computing: A Review*, pages 981–995. Springer Berlin Heidelberg, 2005.

[167] Misha Teplitskiy, Eamon Duede, Michael Menietti, and Karim R. Lakhani. How status of research papers affects the way they are read and cited. *Research Policy*, 51(4):104484, 2022.

[168] Mohammad Tubishat, Mohammad A. M. Abushariah, Norisma Idris, and Ibrahim Aljarah. Improved whale optimization algorithm for feature selection in Arabic sentiment analysis. *Applied Intelligence*, 49(5):1688 – 1707, 2019. Cited by: 152.

[169] Mohammad Tubishat, Norisma Idris, Liyana Shuib, Mohammad A.M. Abushariah, and Seyedali Mirjalili. Improved salp swarm algorithm based on opposition based learning and novel local search algorithm for feature selection. *Expert Systems with Applications*, 145:113122, 2020.

[170] Panagiotis Tzirakis, Jiehao Zhang, and Bjorn W. Schuller. End-to-end speech emotion recognition using deep neural networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5089–5093, 2018.

[171] Lucia Vadicamo, Fabio Carrara, Andrea Cimino, Stefano Cresci, Felice Dell'Orletta, Fabrizio Falchi, and Maurizio Tesconi. Cross-media learning for image sentiment analysis in the wild. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 308–317, 2017.

[172] Hongwei Wang, Fuzheng Zhang, Min Hou, Xing Xie, Minyi Guo, and Qi Liu. SHINE: Signed heterogeneous information network embedding for sentiment link prediction. volume 2018-Febuary, page 592 – 600, 2018. Cited by: 250; All Open Access, Green Open Access.

[173] Kai Wang, Xiaojiang Peng, Jianfei Yang, Shijian Lu, and Yu Qiao. Suppressing uncertainties for large-scale facial expression recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6896–6905, 2020.

[174] Kai Wang, Xiaojiang Peng, Jianfei Yang, Debin Meng, and Yu Qiao. Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Transactions on Image Processing*, 29:4057–4069, 2020.

[175] Kai Wang, Weizhou Shen, Yunyi Yang, Xiaojun Quan, and Rui Wang. Relational graph attention network for aspect-based sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3229–3238, Online, July 2020. Association for Computational Linguistics.

[176] Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. page 7216 – 7223, 2019. Cited by: 162.

[177] Yongjin Wang and Ling Guan. Recognizing human emotional state from audiovisual signals*. *IEEE Transactions on Multimedia*, 10(5):936–946, 2008.

[178] Hajime Watanabe, Mondher Bouazizi, and Tomoaki Ohtsuki. Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE Access*, 6:13825 – 13835, 2018. Cited by: 173; All Open Access, Gold Open Access.

[179] Robert West, Hristo Paskov, Jure Leskovec, and Christopher Potts. Exploiting social network structure for person-to-person sentiment analysis. *Transactions of the Association for Computational Linguistics*, 2, 09 2014.

[180] Guixian Xu, Yueting Meng, Xiaoyu Qiu, Ziheng Yu, and Xu Wu. Sentiment analysis of comment texts based on BiLSTM. *IEEE Access*, 7:51522–51532, 2019.

[181] Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. Double embeddings and CNN-based sequence labeling for aspect extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 592–598, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[182] Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. BERT post-training for review reading comprehension and aspect-based sentiment analysis. volume 1, page 2324 – 2335, 2019. Cited by: 277.

[183] Wei Xue and Tao Li. Aspect based sentiment analysis with gated convolutional networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2514–2523, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[184] Ashima Yadav and Dinesh Kumar Vishwakarma. Sentiment analysis using deep learning architectures: a review. *Artificial Intelligence Review*, 53(6):4335 – 4385, 2020. Cited by: 283.

[185] Wen-Jing Yan, Xiaobai Li, Su-Jing Wang, Guoying Zhao, Yong-Jin Liu, Yu-Hsin Chen, and Xiaolan Fu. CASME II: An improved spontaneous micro-expression database and the baseline evaluation. *PloS one*, 9:e86041, 01 2014.

[186] Huiyuan Yang, Umur Ciftci, and Lijun Yin. Facial expression recognition by de-expression residue learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2168–2177, 2018.

[187] Li Yang, Ying Li, Jin Wang, and R. Simon Sherratt. Sentiment analysis for e-commerce product reviews in Chinese based on sentiment lexicon and deep learning. *IEEE Access*, 8:23522–23530, 2020.

[188] Yilong Yang, Qingfeng Wu, Ming Qiu, Yingdong Wang, and Xiaowei Chen. Emotion recognition from multi-channel EEG through parallel convolutional recurrent neural network. volume 2018-July, 2018. Cited by: 176.

[189] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. XLNet: Generalized autoregressive pretraining for language understanding. volume 32, 2019. Cited by: 2855.

[190] Lijun Yin, Xiaochen Chen, Yi Sun, Tony Worm, and Michael Reale. A high-resolution 3D dynamic facial expression database. In *2008 8th IEEE International Conference on Automatic Face Gesture Recognition*, pages 1–6, 2008.

[191] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and M.J. Rosato. A 3D facial expression database for facial behavior research. In *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, pages 211–216, 2006.

[192] Lin Yue, Weitong Chen, Xue Li, Wanli Zuo, and Minghao Yin. A survey of sentiment analysis in social media. *Knowledge and Information Systems*, 60(2):617 – 663, 2019. Cited by: 195.

[193] Amir Zadeh, Paul Pu Liang, Jonathan Vanbriesen, Soujanya Poria, Edmund Tong, Erik Cambria, Minghai Chen, and Louis-Philippe Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. volume 1, page 2236 – 2246, 2018. Cited by: 345.

[194] Amir Zadeh, Prateek Vij, Paul Pu Liang, Erik Cambria, Soujanya Poria, and Louis-Philippe Morency. Multi-attention recurrent network for human communication comprehension. page 5642 – 5649, 2018. Cited by: 196.

[195] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88, 2016.

[196] Salud María Zafra, Maria Martín-Valdivia, Eugenio Martínez-Cámara, and L. López. Combining resources to improve unsupervised sentiment analysis at aspect-level. *Journal of Information Science*, 42:213–229, 04 2016.

[197] Nianyin Zeng, Hong Zhang, Baoye Song, Weibo Liu, Yurong Li, and Abdullah M. Dobaie. Facial expression recognition via learning deep sparse autoencoders. *Neurocomputing*, 273:643 – 649, 2018. Cited by: 392.

[198] Sara Zhalehpour, Onur Onder, Zahid Akhtar, and Cigdem Eroglu Erdem. BAUM-1: A spontaneous audio-visual face database of affective and mental states. *IEEE Transactions on Affective Computing*, 8(3):300–313, 2017.

[199] Chen Zhang, Qiuchi Li, and Dawei Song. Aspect-based sentiment classification with aspect-specific graph convolutional networks. page 4568 – 4578, 2019. Cited by: 254.

[200] Feifei Zhang, Tianzhu Zhang, Qirong Mao, and Changsheng Xu. Joint pose and expression modeling for facial expression recognition. page 3359 – 3368, 2018. Cited by: 174.

[201] Jianhua Zhang, Zhong Yin, Peng Chen, and Stefano Nichele. Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review. *Information Fusion*, 59:103 – 126, 2020. Cited by: 212.

[202] Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126(10):1084 – 1102, 2018. Cited by: 307; All Open Access, Green Open Access.

[203] Shiqing Zhang, Shiliang Zhang, Tiejun Huang, and Wen Gao. Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching. *IEEE Transactions on Multimedia*, 20(6):1576–1590, 2018.

[204] Shiqing Zhang, Shiliang Zhang, Tiejun Huang, Wen Gao, and Qi Tian. Learning affective features with a hybrid deep model for audio-visual emotion recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10):3030 – 3043, 2018. Cited by: 183.

[205] Shunxiang Zhang, Zhongliang Wei, Yin Wang, and Tao Liao. Sentiment analysis of Chinese micro-blog text based on extended sentiment dictionary. *Future Generation Computer Systems*, 81:395–403, 2018.

[206] Tong Zhang, Wenming Zheng, Zhen Cui, Yuan Zong, and Yang Li. Spatial–temporal recurrent neural network for emotion recognition. *IEEE Transactions on Cybernetics*, 49(3):839–847, 2019.

[207] Xing Zhang, Lijun Yin, Jeffrey F. Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M. Girard. BP4D-spontaneous: a high-resolution spontaneous 3D dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014. Best of Automatic Face and Gesture Recognition 2013.

[208] Zheng Zhang, Jeffrey M. Girard, Yue Wu, Xing Zhang, Peng Liu, Umur Ciftci, Shaun Canavan, Michael Reale, Andrew Horowitz, Huiyuan Yang, Jeffrey F. Cohn, Qiang Ji, and Lijun Yin. Multimodal spontaneous emotion corpus for human behavior analysis. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3438–3446, 2016.

[209] Guoying Zhao, Xiaohua Huang, Matti Taini, Stan Z. Li, and Matti Pietikäinen. Facial expression recognition from near-infrared videos. *Image and Vision Computing*, 29(9):607–619, 2011.

[210] Han Zhao, Shanghang Zhang, Guanhang Wu, João P. Costeira, José M.F. Moura, and Geoffrey J. Gordon. Adversarial multiple source domain adaptation. volume 2018-December, page 8559 – 8570, 2018. Cited by: 222.

[211] Jianfeng Zhao, Xia Mao, and Lijiang Chen. Speech emotion recognition using deep 1D 2D CNN LSTM networks. *Biomedical Signal Processing and Control*, 47:312–323, 2019.

[212] Yuxin Zhao, Sixiang Cheng, Xiaoyan Yu, and Huilan Xu. Chinese public's attention to the COVID-19 epidemic on social media: Observational descriptive study. *Journal of Medical Internet Research*, 22(5), 2020. Cited by: 165; All Open Access, Gold Open Access, Green Open Access.

[213] Wei-Long Zheng, Wei Liu, Yifei Lu, Bao-Liang Lu, and Andrzej Cichocki. EmotionMeter: A multimodal framework for recognizing human emotions. *IEEE Transactions on Cybernetics*, 49(3):1110–1122, 2019.

[214] Wei-Long Zheng and Bao-Liang Lu. Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. *IEEE Transactions on Autonomous Mental Development*, 7(3):162–175, 2015.

[215] Wei-Long Zheng, Jia-Yi Zhu, and Bao-Liang Lu. Identifying stable patterns over time for emotion recognition from EEG. *IEEE Transactions on Affective Computing*, 10(3):417–429, 2019.

[216] Nazan Öztürk and Serkan Ayvaz. Sentiment analysis on twitter: A text mining approach to the Syrian refugee crisis. *Telematics and Informatics*, 35(1):136–147, 2018.