392167 Automatic Human Behaviour Analysis Using Machine Learning and Psychological Methods

Project Report: Automatic Detection of Comprehension Problems during Online Lectures

Author:

Module: 39-M-Inf-P Projekt

Ungraded

https://github.com/maskaljunas/ComprehensionProblems

1 Introduction

Due to the COVID-19 pandemic students and teachers all over the world are faced with online teaching. With not being present in the lecture room teachers cannot keep track of all students reactions in a virtual classroom. In real life classrooms, even with a high number of participants, it is easy to spot students sending signals of engagement as boredom or confusion and respond appropriately to it, e.g. by giving an easier explanation. In an online classroom, however, looking for reactions of students is made difficult due to various reasons. Only a small portion of students is directly visible on the screen, for seeing the others one have to select manually the screen of more students. Hence when monitoring the engagement of the students the teacher will be distracted and will not be able to give a structured and understandable lecture.

Unlike, not monitoring the engagement of students might also lead to dissatisfaction for both sides when students are left alone with their problems in understanding the subject matter.

Since these situations are not familiar to most of the students, they are restrained in behaving naturally when being recorded during a lecture. Often their appearance on the learning platform is more interesting and distracting than following the lesson. Furthermore, the absence of other students on the room might also affect the individual's self-confidence in a negative way such that the courage to ask questions is not present. Combining those factors, online teaching can be frustrating for both, students and teachers.

2 Theory and Related Work

2.1 Theory of Comprehension of Complex Material

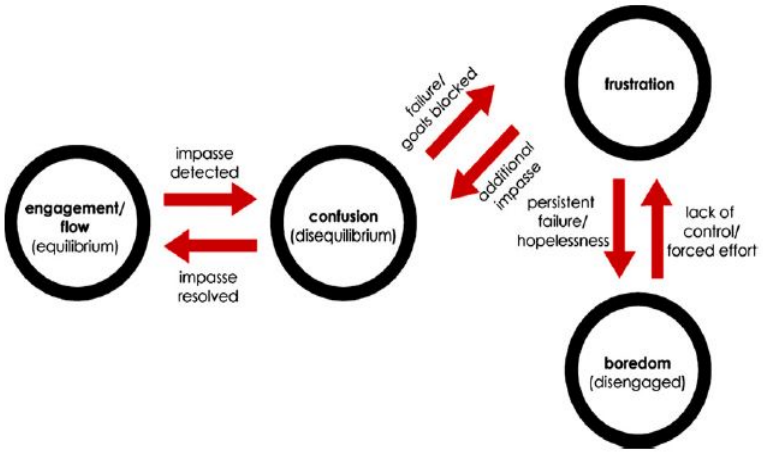
The term comprehension is defined as the "ability to understand completely and be familiar with a situation, facts" [7]. Usually, it refers to a specific type of skill as listening or reading comprehension. While the second presupposes listening comprehension, the latter is developed during childhood and developmental problems can cause general language comprehension deficits. Deafness, autism spectrum disorder or other neurological impairments affect language comprehension on different levels. For reading comprehension, the simple word reading activates different cognitive processes than the comprehension of complex facts and hence does not ensure good comprehension skills

[6][24]. Unless there are no neurological impairments these two skills are necessary for following a lecture and comprehension problems during learning are caused by different reasons.

As proposed by Graesser and D'Mello, learning and understanding complex material is an effortful deep level cognitive process of "reflection and inquiry because there is a discrepancy between (a) the immediate situation and (b) the person's knowledge, skills, and strategies". This state of an unaccomplished merging of (a) and (b), to be more precise contradictions, incongruities, uncertainty and obstacles to goals, is defined as the cognitive disequilibrium. Once the cognitive equilibrium is restored due to conquered challenges a state of flow arises [15]. The concept of flow is introduced by Csikszentmihalyi as an optimum level of concentration and engagement, a complete immersion in an activity when completing a task [9]. In the context of learning the learning material, its structure, segmentation and the learning strategy are congruent with the student's comfort and learning potential [5][15][28].

To illustrate the interaction and dependence of different emotional stages of the cognitive disequilibrium framework figure 1 is taken from Graesser and D'Mello [15]. Starting in an engagement or flow state, which is defined as cognitive equilibrium, it can be shifted towards a cognitive disequilibrium by an event or thought towards the affective state of confusion. This state can be immediately resolved back to the state of equilibrium or if the challenges are not negotiated the affective state changes into frustration. Additional difficulties such as persistent failure result in disengagement from the task and in the end in boredom. From each state, an oscillation of affective states is possible following the links of the transitions. Hence comprehension problems in learning environments are caused by an imbalance of the student's current knowledge and their to be achieved knowledge objectives being not resolved.

Figure 1: Cognitive disequilibrium framework. Taken without permission from Graesser and D'Mello [15].



2.2 What Emotions occur during Learning?

Graesser and D'Mello evaluated in several studies emotions that are elicited during learning of difficult material with AutoTutor and ARIES. The AutoTutor is an intelligent tutoring system covering topics that enable students to learn critical thinking and other skills by monitoring their dialogue and facing them with questions that require reasoning. ARIES is a similar system teaching scientific critical thinking [15]. Confusion, frustration, boredom, flow/engagement and neutral affective states were observed by trained judges during an interaction with the AutoTutor of 34 students. Those findings are supported by their other studies [10][11][15].

Baker et al. as well reported confusion, frustration, boredom, flow/engagement, delight, surprise and neutral affective states during three different computerized learning environments [2].

Affective states in massive open online courses (MOOCs) were investigated by Xiao et al. showing engagement, boredom, confusion, frustration, delight, surprise and curiosity in 22 students during a passive learning scenario. Moreover, they focused on affective states transitions as postulated by Graesser and D'Mello [15]. Significant transitions from engagement to boredom and from confusion to engagement were present, indicating that passive learning scenarios like massive lectures might be a determinant for early entering the state of boredom than in direct interaction learning [29].

Another study showed that in the majority of cases neutral, angry and sad emotions were present during a computer-based assessment. The emotions of 172 students were determined by two experts and the FaceReader, an application developed by Vicar Vision and Noldus Information Technology bv that is able to recognize human emotions. Their ratings were compared and revealed an overall agreement of 87% [27].

In recent years the focus shifted toward the automatic detection of affect, especially in improving the classification of different affective states. Automatic detection of engaged, bored and neutral states by using convolutional networks reveal average accuracy scores of 86% for posed classroom affects and 70% for spontaneous classroom affects [1].

Gupta et al. provided a data set (DAISEE) with videos for affect recognition containing real-world settings of students participating in a MOOC. The videos are labelled according to the students' concurrent state of engagement, boredom, confusion and frustration allowing a multiclass classification with accuracies of 51.07%, 35.89%, 57.45% and 73.09%, respectively. On closer consideration, they noticed complementary behaviour of engagement and boredom, when both are not present or very low confusion or frustration dominate. It shows that the transition and interaction of affective states in such environments have to be investigated more closely such that the insights of their dynamic improve affect recognition in general [20].

In any case it holds that affective states do depend on the learning environment. Students in classrooms express their emotions differently than as passive listeners in lectures with low or no interaction with the teacher since there is no purpose of communication. Single student learning scenarios with tutoring systems when dealing with complex material, in turn, do show other behaviour of students. When analysing and interpreting affect of students their environmental context must be taken into account however key expressions such as engagement, confusion, frustration and boredom can indicate their learning success or failure.

2.3 How do emotions affect learning?

The distinction of emotions in positive and negative dimensions as proposed by Greenwald et al. [18] and their impact on cognitive processes was studied in an EEG experiment. A discrimination task was performed after subjects' emotions were manipulated into positive, negative and neutral states. Results show that in the negative emotional condition the P300 component, an event-

related potential indicating that a target is detected was smaller than in other conditions. These findings suggest that negative emotions may use the resources of task-related attention processes and thus diminish cognitive capacity during learning [23].

Positive emotions such as happiness and pride, by contrast, correlate with the interest and motivation of students leading to the achievement of flow, an optimum state of engagement into a task [25]. A behavioural study in 1979 by Masters et al. already showed similar results for preschool children that positive affective states enhance interest, involvement and arousal whereas negative emotions decreased them [22]. This simple dichotomy however does not hold as a general rule.

The state of confusion, occurring when there exists an incongruence of current knowledge and an external event like transfer to new awareness, can lead, if persistent, to frustration and disengagement. However, confusion has a more important role for learning than simply leading to further frustration. In the emotion evaluation study of D'Mello and Graesser mentioned above, a pretest, subsequent interaction with the AutoTutor and a posttest in terms of measurement for knowledge improvement were made.

A correlation analysis between learning gains, a measurement for knowledge improvement, confusion and flow/engagement showed a positive correlation (r=.33; r=.29) but negative for boredom (r=-.39). They suggest that confusion is a good predictor for learning improvement and hence cognitive disequilibrium plays a key role in learning and comprehension [8][15]. Graesser and D'Mello argue that confusion acts as the gateway for entering either the flow level or frustration state and should be paid more attention regarding the regulation of emotions during a learning scenario [15].

2.4 Semantic Features corresponding to learning or CP

Detection of affective states can be performed through various channels of communication (e.g., dialogue, facial expressions, body language). Facial expressions convey meaning in the sense that they reflect one's inner emotions, as long as they are not suppressed on purpose. As Ekman and Friesen proposed there exist six basic emotions (fear, disgust, happiness, sadness, surprise and anger) that are expressed the same way regardless of the person's cultural background [12]. However higher cognitive processes such as reasoning, e.g. during learning, elicit more complex expressed emotions and hence a universal generalisation of their expression should not be expected.

To identify those emotions, the facial coding system (FACS) was developed by Ekman and Friesen [14]. The coding system uses single or combined muscle contractions as an Action Unit and assigns them to corresponding expressed emotions. Key emotions playing a relevant role listed above are defined as follows:

Confusion is shown by a lowered brow (AU4), tightening of the eyelids (AU7) and an absence of lip corner pull (AU12) [16][15]. Other studies reported the frequency of mouth dimpling (AU14) during thoughtful and concentrated tasks being positively correlated with learning gains after the state of confusion [21][16].

Borges et al. suggest that the expression of confusion and boredom is not the simultaneous presence of a combination of AUs but rather a subset of those over consecutive time points [4]. The systematic description of boredom is not consistent with specific action units, rather it is an expressionless face which is difficult to distinguish from neutral face [15].

For frustration inner and outer brow-raising (AU1, AU2) were reported by Graeser and D'Mello, however, Graafsgaard et al. showed a positive correlation of the lowered brow (AU4) [15][16]. In another study, Graafsgaard et al. also reported that mouth dimpling (AU14) was an indicator for frustration and lower learning gain unlike as stated before [16][21] . Additionally, postural movement increased with higher frustration as well as near postural position to the display [17].

Studies including body posture to detect affect during interacting with tutoring systems used sensors, such as pressure sensors, such that

3 Approach

3.1 Data

A supervised learning approach, in this case classification, was used to build a comprehension problem detection model. For this, the DAiSEE (Dataset for Affective States in E-Environments) data set was used. It was collected to recognise user engagement in the wild by Gupta et al. in 2016 [19]. The data set contains 9.068 frontal web camera videos of 112 Asian students (m=80; f=32; age=18-30) watching two different stimuli of 20 minutes on their notebook. The video data is then cut in video snippets of 10 seconds containing 300 frames and was labelled according to the affective state of boredom, confusion, frustration and engagement, all of the labels having four

levels (very low, low, high, very high). All crowd annotations were correlated with annotations of psychologists experts to ensure overall agreement. The data has a an original 20:60:20 split which was kept.

3.2 Data Labelling

For the purpose of classifying comprehension problems, the original labels were modified without changing their meaning. As defined in section 2.1, the affective states confusion and frustration contribute to comprehension problems and hence are the key labels. The first labelling cluster defines comprehension problems as those videos where the original label is stated as confusion or frustration with a higher value than zero regardless of the intensity for boredom and engagement. This cluster results in a balanced data set. Another cluster of labels is added with a higher threshold of greater than or equal to two for confusion or frustration, resulting in an unbalanced data set (see Table 1). However, this option might be more suitable, since by looking through the videos it was hard to comprehend their given labels. Since the participants are from Indian origin and there is evidence for the importance of cultural familiarity regarding facial emotion recognition it might contribute to this misinterpretation from my point of view [13]. Nevertheless those not clearly expressed facial expressions also might be problematic for training a classification model.
Additionally, boredom was added after the model's performance was evaluated. Due to low classification results it was assumed this option might reveal more information. For overview of the labelling cluster have a look at Table 1. These five variations of labelling clusters were used for selecting the best classification model.

Boredom	Engagement	Confusion	Frustration	BorEng=0 ‖ CP=1
>0	>0	0	0	0
x	x	≥ 1	≥ 1	1
Boredom	Engagement	Confusion	Frustration	BorEngLess=0 ‖ CP=1
>1	>1	0	0	0
x	x	≥ 2	≥ 2	1
Boredom	Engagement	Confusion	Frustration	BorEngFrus=0 ‖ onlyConfused=1
x	x	0	x	0
x	x	≥ 1	x	1
Boredom	Engagement	Confusion	Frustration	BorEngCon=0 ‖ onlyFrustrated=1
x	x	x	0	0
x	x	x	≥ 1	1
Boredom	Engagement	Confusion	Frustration	EngConFrus=0 ‖ onlyBoredom=1
0	x	x	x	0
≥ 1	x	x	x	1

Table 1: Five labelling clusters. CP is defined as confusion or frustration.

3.3 Feature Extraction

For feature extraction, the toolkit OpenFace was used to extract relevant facial and upper body behaviour. OpenFace, an open-source framework, based on deep neural network algorithms estimates head pose and eye gaze, detects facial landmarks and recognises facial action units [3]. It is using its input files for extracting the requested facial features by analysing them frame by frame. For this data set each video file results in 300 values for the respective feature. Since a frame by frame analysis does not include facial and upper body changes but rather reflects only one single point in time during a sequence of dynamically expressed emotions, time is a key factor to extract changes and most salient emotions. Hence, various parameters as the arithmetic mean, standard deviation and other measured parameters are used, depending on the feature property.

18 Action Units can be extracted with OpenFace (see Table 2). Each video frame is classified in a present or not present AU (0 or 1) and indicates its intensity in a range of 1-5. Only the presence of the respective AU was used for further analysis. For all AU the average of 300 frames, representing the whole 10 seconds, was calculating. The value indicates how often or how long a certain Action Unit occurs also being already normalised for feature comparison. Nevertheless, the sum of Action Units was calculated. Since

AU45, a blink detector, the presence of a relaxed upper eyelid muscle in each frame rather represents how many frames the eyelids were closed than how many blinks were present, blink count was calculated separately (see feature_extraction.py). Inter blink interval (in frames) and the average AU45 frame length were also calculated being a potential indicator for tiredness and hence boredom [26].

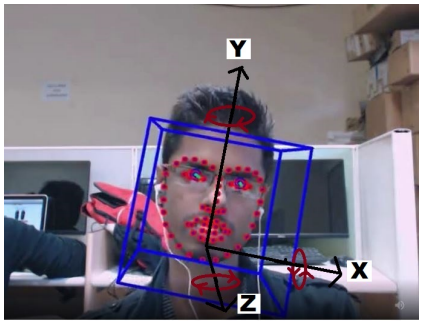
AU	Description	AU	Description	AU	Description
AU1	Inner brow raiser	AU9	Nose wrinkler	AU20	Lip stretched
AU2	Outer brow raiser	AU10	Upper lip raiser	AU23	Lip tighter
AU4	Brow lowerer	AU12	Lip corner puller	AU25	Lips apart
AU5	Upper lid raiser	AU14	Dimpler	AU26	Jaw drop
AU6	Cheek raiser	AU15	Lip corner depressor	AU28	Lip suck
AU7	Lid tighter	AU17	Chin raiser	AU45	Blink

Table 2: Facial Action Units extracted by OpenFace.

Mouth opening activity was calculated based on AU25 on the same procedure as for blink count. Counts of how many times the mouth was opened, inter mouth open interval and its average frame length of the mouth being open ware included into the features. Mouth opening activity reflects heavy breathing or laughing behaviour, which might contribute to behaviour explanation.

Eye gaze direction change of both eyes for left-right and up-down gaze were used to calculate the standard deviation of all frames. Left-right gaze movement is represented by a value change from positive to negative, up-down gaze in negative to positive change. The standard deviation, in this case, might indicate how often and how much gaze movement changed. A stable gaze straight ahead would result in values around zero, as well as its standard deviation. For gaze fixations, the difference of consecutive gaze direction change was computed and values of around zero counted, suggesting low or no gaze change. For both movements directions, the standard deviation of the differences were computed. Average head distance to the camera, as well as its maximum and

Figure 2: Head axis for illustration.



minimum values, were calculated. Head rotation around yaw (left-right rotation) and roll (left-right head skew) dimensions were also taken into account (see Figure 2). The average and maximum rotation for every side (left-right) were calculated. In total 59 features were calculated.

3.4 Behavioural/ Feature Data Preprocessing

All features were viewed according to their distribution by hand. To ensure their equal contribution to the model, the features were normalized and in case of skewness a log transformation was applied, also reducing the effects of outliers. Aside from the named changes, another two methods were applied: Outliers were detected and either capped by replacing them with their 2% and 98% quantile or simply dropped. All three versions were used for model training.

To check for feature independence a correlation plot was generated and due to the high linear relationship of the arithmetic mean and the sum of AUs, the sums were dropped (see Figure 8 in appendix). Moreover, AU1 and AU2 were merged because they show a higher linear relationship as well as they were reported appearing together and the respective original AUs were dropped. All features were inspected according to their variance however no more features were excluded by hand due to further exploration and hence resulting in 40 different features.

Recursive feature elimination with cross-validation was applied to inspect the most promising features. With oversampling the minority class in every case to improve the class distributions and iterating about forty times to get robust data, it yields different numbers of features for different labelling clusters. However, consistent features were always present in the best selected features list. Comparing the classification performance of the different labelling clusters with a logistic regression classifier, it shows that the first labelling cluster with comprehension problems defined as confusion and frustration higher than zero gives the best results (see Figure 9 and 10 in appendix). Based on that, fourteen features were selected for further model comparison (see Table 4 in appendix). Nevertheless, the high variation of the best number of features up to 31 already shows the poor classification performance of the model. Clusters with capped and dropped data also decrease in their classification performance. Due to the high elimination of the target label in capped and dropped data, it is assumed those outliers might be important for detecting the target (for class distribution see Table 5 in

appendix).

3.5 Model Evaluation

To select the best classifier (Logistic Regression, K-Nearest Neighbor, Support Vector Classification, Random Forest and Decision Tree) each model was trained on the data with resulting 14 features and then their hyper-parameters were optimized. Since the distribution of the labels is slightly unbalanced the minority class, in this case data labelled as comprehension problem, was over-sampled to reduce the non-target bias. Accuracy rates were not the most appropriate metrics to look for and hence Precision-Recall ratio as well as ROC-AUC were taken as primary metric when evaluating on the test set (see Figure 5). They show minor improvements between the models, however a cross-validation reveal slightly better results for the Random Forest classifier (Figure 3). Nevertheless the results show a poor model performance. Random Forest classifier model was the trained on the whole data set for the application in tutoring system. For overall recognition results see Table 3.

Figure 3: Classifier comparison of 10-fold cross-validation.

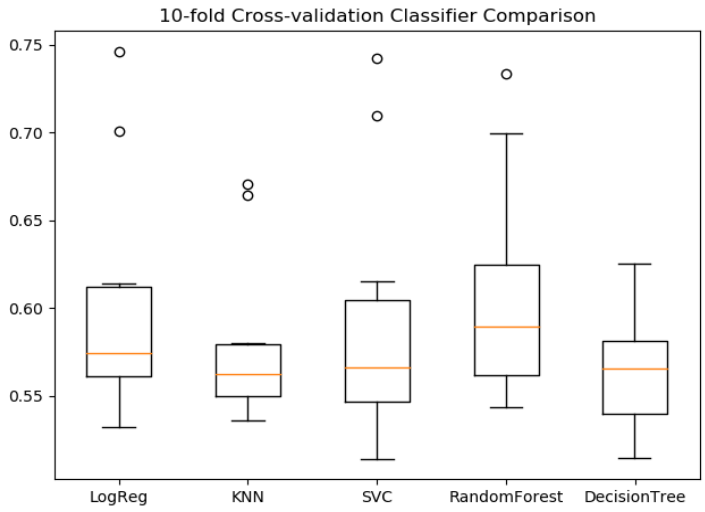
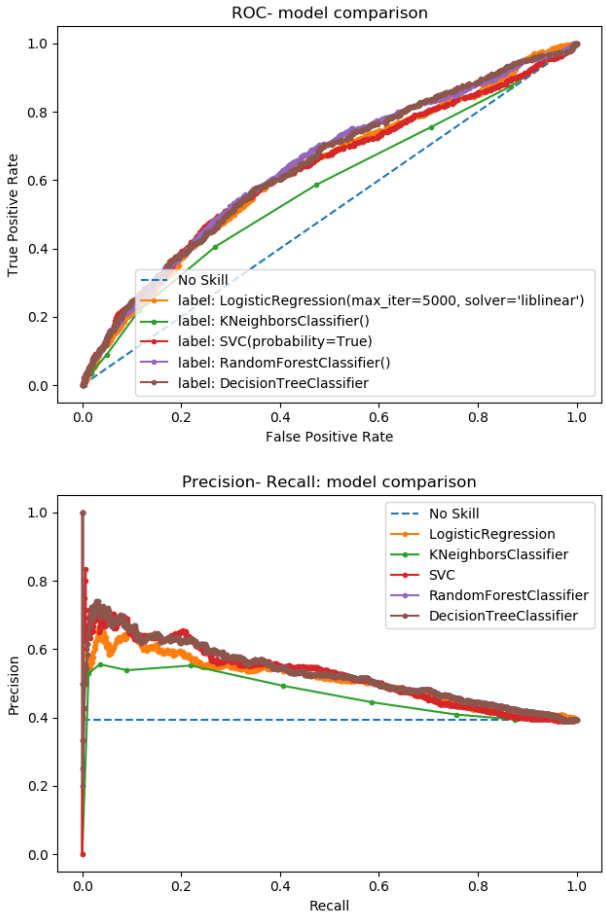


Figure 5: ROC-curve and Precision-Recall-curve comparison of selected classifiers.



4 User Study

To test the models' practicability for new data and inspect its contribution detecting comprehension problems during learning scenarios, an intelligent tutoring (ITS) system was developed. The tutoring system, while presenting complex material to comprehend and apply at the end, will detect comprehension problems and give an option to see some of the material again. This

Model	Accuracy	F-1 Score	ROC-AUC	Precision-Recall
Dummy Baseline	0.522	0.603		
Logistic Regression	0.626	0.256	0.637	0.512
K-Nearest Neighbour	0.607	0.327	0.584	0.458
Support Vector C.	0.617	0.182		
Random Forest	0.639	0.422	0.644	0.527
Decision Tree	0.5812	0.400	0.528	0.431

Table 3: Test set evaluation metrics on different classifiers.

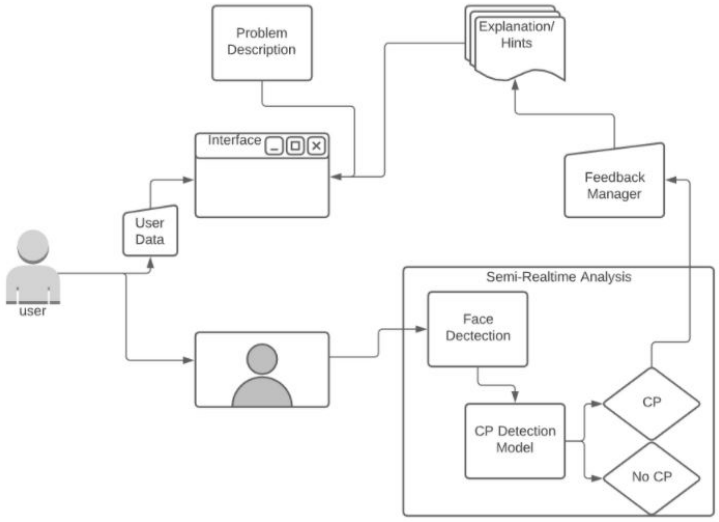
should, if correctly detected, help the user solving the task. For this a quantitative cross-sectional study with randomised controlled trials was designed. A control group was interacting with the tutoring system but does not have the option to get further explanation, even if comprehension problems are detected. In the experimental group the system, however, will adapt to the subjects' needs and present an overview of the material, if desired. To defend the comprehension problem detection model, even if based on poor performance, the experimental group and the control group should have a different amount of correct answers in the task. Furthermore they also might have distinct response times, since the experimental group has the option to get an overview of the presented material and hence the information will be repeated and be retrieved faster from the memory.

4.1 Intelligent Tutoring System

An intelligent tutoring system uses principles of artificial intelligence such as knowledge representation, inference mechanisms and machine learning for operating. In this case, a machine learning based classification algorithm is used to detect comprehension problems of users interacting with the tutoring system. The tutoring system presents learning material of an explanation of how to form the basic Finnish plural. In the meanwhile, a webcam is recording the subject's upper body for further data processing. As done in the DAISEE data set, the video stream consist of 10 seconds chunks (300 frames) to ensure similar feature modality. A face detection model examines each frame and forwards it for feature extraction with openFace. 14 features are calculated as elaborated in the comprehension problems detection model and then the comprehension problem detection model is applied. The model's predictions are stored and verified for detected comprehension problems. Due to the high computational effort of feature extraction, their calculation and

preprocessing, a delay of about 10 to 18 seconds results. However, it is overcome by a break between the end of the explanation and the task or further explanation. If the predictions contain at least one detected comprehension problem, then the subject has the option to be faced with an overview of the presented material for 30 seconds. After that, a task of forming eight plural words must be completed. The answers as well as the response times were recorded for further analysis. Once the task is done, the system closes. Figure 6 displays an overview of tutoring system's architecture.

Figure 6: General architecture of the developed intelligent tutoring system.



4.2 Stimuli

To present the subjects with complex material in order to induce comprehension problems during an experiment, the stimuli must correspond an adequate level of complexity and minor time investment. New language acquisition does fulfil these properties. Grammar explanations are complex enough and can be executed within a short time. For this, the basic Finnish plural form suits very well, because there exist three exceptions within a basic rule.

Figure 7: Overview of the Finnish plural form. This is the stimulus subjects can have a look at if comprehension problems are detected. The whole stimulus ca be seen in the project's repository at GitHub.

4.3 Study Report

Due to the Corona crisis, the planned number of participants for the study could not be realised. An idea to conduct this as an online study was neglected because of practical and privacy reasons. Only six participants took part in the study (N-controll=3; N-experimental=3). All participant were properly informed and instructed, such that lightning conditions and body position relative to the camera were optimal. Information about data handling and storing was given before the experiment was conducted and all participants signed an informed consent form (see appendix).

Eight plural words, which had to be formed in the task were summed up by their total number of letters and mistakes were subtracted divided by total amount of letter, resulting in an accuracy score. The control group exhibit a mean accuracy of 0.945 (sd=0.014) and the experimental group of 0.933 (sd=0.38). For the sake of completeness an unpaired t-test was conducted with, surprisingly, no significant results ($t(4)=0.3536$, $p=0.074$). No more test, such as variance comparison, were executed. In the experimental group the model detected only in one subject comprehension problems, however during observation of the experiment based on my personal judgement at

least another subject showed expressions of comprehension problems such as confusion. An oral interview of the subject with detected comprehension problems reveal indeed that the repetition of the explanation helped to solve the task faster. The subject reported to remember and reflect the rules to apply after the hint.

5 Discussion and Conclusion

Detection of comprehension problems during learning scenarios is a crucial component for developing intelligent tutoring systems that adapt to the user's behaviour and learning progress. Based on the affect interactions during learning complex material as proposed by Graesser and D'Mello it is important to detect confusion and frustration as the main affective cognitive states, which when not resolved lead to disengagement [15]. This project had the aim to develop such a model that is able to detect those cognitive states in order to react appropriate. In this approach confusion and frustration was taken as a single state, exposing similar behavioural cues. Results from section 3.5 illustrate the poor performance of the model using this approach. Confusion as well as frustration are complex emotional expressions which cannot be reduced into single or combined Action Units. Rather their expression is a sequence of consecutive different behavior over time and hence a frame by frame analysis yields such dissatisfied detection results. Furthermore their expression depends on the context and is not a sequence of same behavior. Therefore, those two cognitive states and their expressions must be explored in further research. Indeed the approach of examining a time interval rather than single frames is a step into the right direction, however the selected interval of 10 seconds might be too long. Emotional expressions can be expressed within milliseconds but also seconds, however they usually not not last for long and rather several affective states overlap during the interval, especially if their stimulus has a lot information to process. As seen in the data set the original labelling shows overlapping affective states. Confusion and frustration appear together with engagement or boredom, not knowing if the subjects showed this expression in the same time or at different time points during the 10 seconds. Only Engagement and boredom reveal contrary relation. A detailed look into the distribution of the classes and their data showed bi-modal distribution. The question arises whether there exist different types of groups showing distinct expressions while ex-

periencing the same affective state. Nevertheless the developed model failed to detect comprehension problems. In this context the major problem is the trade off between False Negatives and False Positives. A tutoring system, that does not detect comprehension problems (FN), is facing profound consequences, since it defeats its purpose and is in the end useless. Rather more falsely positive detected situations would results in more sensitive model towards comprehension problems detection. A attempt in using boredom as a hint for comprehension problems, since this state is caused by comprehension problems, revealed similar accuracy results leading to no information gain.

References

[1] TS Ashwin and Ram Mohana Reddy Guddeti. "Automatic detection of students' affective states in classroom environment using hybrid convolutional neural networks". In: *Education and Information Technologies* 25.2 (2020), pp. 1387–1415.

[2] Ryan SJd Baker et al. "Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments". In: *International Journal of Human-Computer Studies* 68.4 (2010), pp. 223–241.

[3] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. "Openface: an open source facial behavior analysis toolkit". In: *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE. 2016, pp. 1–10.

[4] Niklas Borges et al. "Classifying Confusion: Autodetection of Communicative Misunderstandings using Facial Action Units". In: *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. IEEE. 2019, pp. 401–406.

[5] Ann L Brown, Sybil Ellery, and Joseph C Campione. *Creating zones of proximal development electronically.* Lawrence Erlbaum Associates Publishers, 1998.

[6] Kate Cain and Jane Oakhill. *Children's comprehension problems in oral and written language: A cognitive perspective.* Guilford Press, 2008.

[7] *COMPREHENSION — meaning in the Cambridge English Dictionary*. https://dictionary.cambridge.org/dictionary/english/comprehension. (Accessed on 12/10/2020).

[8] Scotty Craig et al. "Affect and learning: an exploratory look into the role of affect in learning with AutoTutor". In: *Journal of educational media* 29.3 (2004), pp. 241–250.

[9] Mihaly Csikszentmihalyi and Mihaly Csikzentmihaly. *Flow: The psychology of optimal experience*. Vol. 1990. Harper & Row New York, 1990.

[10] Sidney K D'Mello, Scotty D Craig, and Art C Graesser. "Multimethod assessment of affective experience and expression during deep learning". In: *International Journal of Learning Technology* 4.3-4 (2009), pp. 165–187.

[11] Sidney K D'Mello et al. "Automatic detection of learner's affect from conversational cues". In: *User modeling and user-adapted interaction* 18.1-2 (2008), pp. 45–80.

[12] Paul Ekman and Wallace V Friesen. "Constants across cultures in the face and emotion." In: *Journal of personality and social psychology* 17.2 (1971), p. 124.

[13] Hillary Anger Elfenbein and Nalini Ambady. "When familiarity breeds accuracy: cultural exposure and facial emotion recognition." In: *Journal of personality and social psychology* 85.2 (2003), p. 276.

[14] E Friesen and Paul Ekman. "Facial action coding system: a technique for the measurement of facial movement". In: *Palo Alto* 3 (1978).

[15] Arthur C Graesser and Sidney D'Mello. "Emotions during the learning of difficult material". In: *Psychology of learning and motivation*. Vol. 57. Elsevier, 2012, pp. 183–225.

[16] Joseph Grafsgaard et al. "Automatically recognizing facial expression: Predicting engagement and frustration". In: *Educational Data Mining 2013*. 2013.

[17] Joseph F Grafsgaard et al. "The additive value of multimodal features for predicting engagement, frustration, and learning during tutoring". In: *Proceedings of the 16th International Conference on Multimodal Interaction*. 2014, pp. 42–49.

[18] Mark K Greenwald, Edwin W Cook, and Peter J Lang. "Affective judgment and psychophysiological response: dimensional covariation in the evaluation of pictorial stimuli." In: *Journal of psychophysiology* (1989).

[19] Abhay Gupta et al. "DAISEE: dataset for affective states in e-learning environments". In: *arXiv* (2016), pp. 1–22.

[20] Abhay Gupta et al. "Daisee: Towards user engagement recognition in the wild". In: *arXiv preprint arXiv:1609.01885* (2016).

[21] Gwen C Littlewort et al. "Automated measurement of children's facial expressions during problem solving tasks". In: *Face and Gesture 2011*. IEEE. 2011, pp. 30–35.

[22] John C Masters, R Christopher Barden, and Martin E Ford. "Affective states, expressive behavior, and learning in children." In: *Journal of Personality and Social Psychology* 37.3 (1979), p. 380.

[23] Jörg Meinhardt and Reinhard Pekrun. "Attentional resource allocation to emotional events: An ERP study". In: *Cognition and Emotion* 17.3 (2003), pp. 477–500.

[24] Scott G Paris et al. "Spurious and genuine correlates of children's reading comprehension". In: *Children's reading comprehension and assessment* (2005), pp. 131–160.

[25] Reinhard Pekrun et al. "Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative research". In: *Educational psychologist* 37.2 (2002), pp. 91–105.

[26] John A Stern, Donna Boyer, and David Schroeder. "Blink rate: a possible measure of fatigue". In: *Human factors* 36.2 (1994), pp. 285–297.

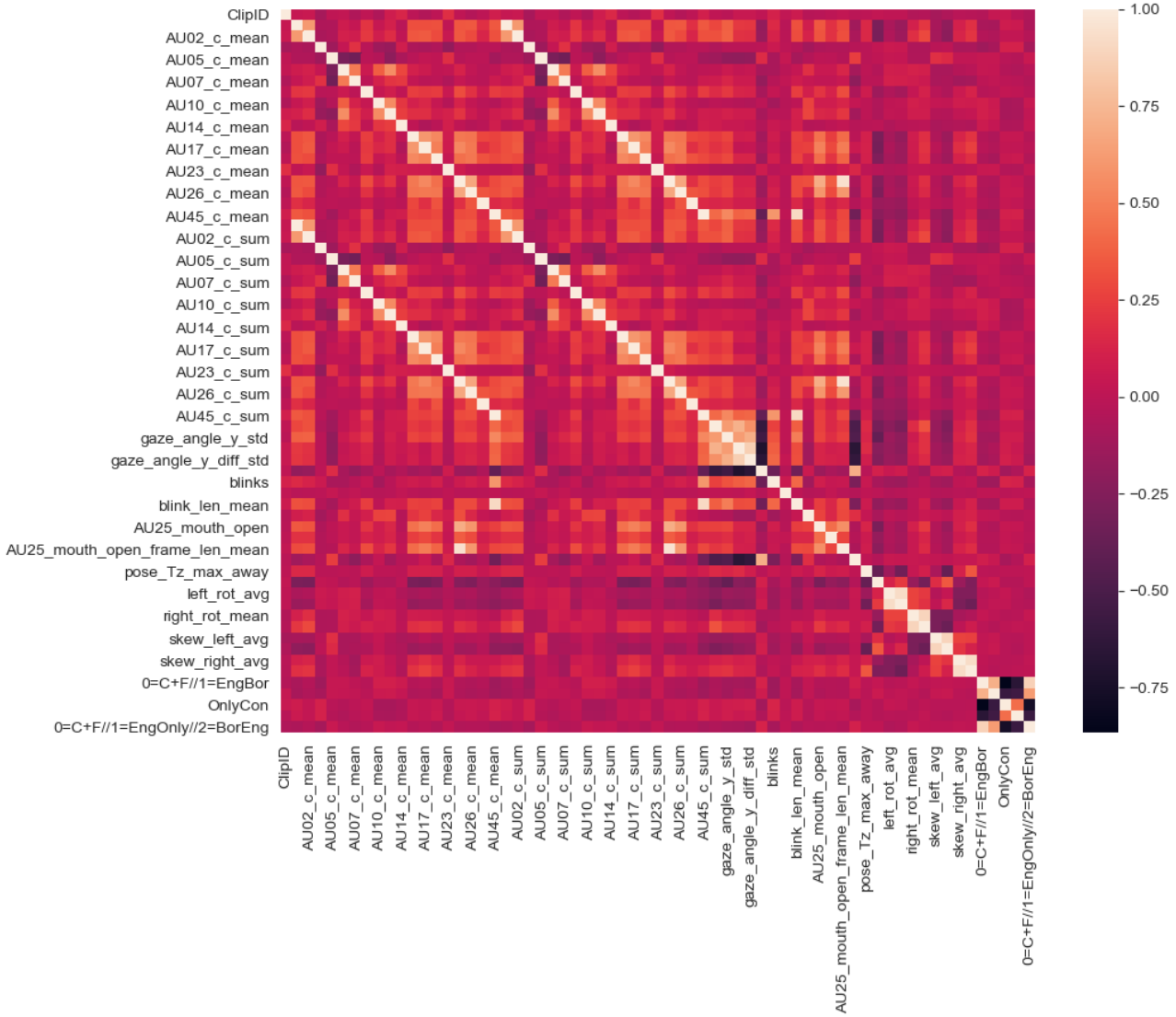
[27] Vasileios Terzis, Christos N Moridis, and Anastasios A Economides. "Measuring instant emotions based on facial expressions during computer-based assessment". In: *Personal and ubiquitous computing* 17.1 (2013), pp. 43–52.

[28] Lev Vygotsky. "Interaction between learning and development". In: *Readings on the development of children* 23.3 (1978), pp. 34–41.

[29] Xiang Xiao, Phuong Pham, and Jingtao Wang. "Dynamics of affective states during mooc learning". In: *International Conference on Artificial Intelligence in Education*. Springer. 2017, pp. 586–589.

A Appendix

Figure 8: Feature correlation plot of original train data.



AU01_c_mean	AU04_c_mean	AU12_c_mean
AU20_c_mean	AU25_c_mean	AU28_c_mean
gaze_angle_x_diff_std	gaze_angle_y_diff_std	gaze_fixation_count
blinks	pose_Tz_max_away	left_rot_mean
right_rot_mean	left_rot_max	

Table 4: 14 best Features evaluated with recursive feature elimination and cross-validation.

Figure 9: ROC-curve of different labelling clusters with logistic regression classifier.

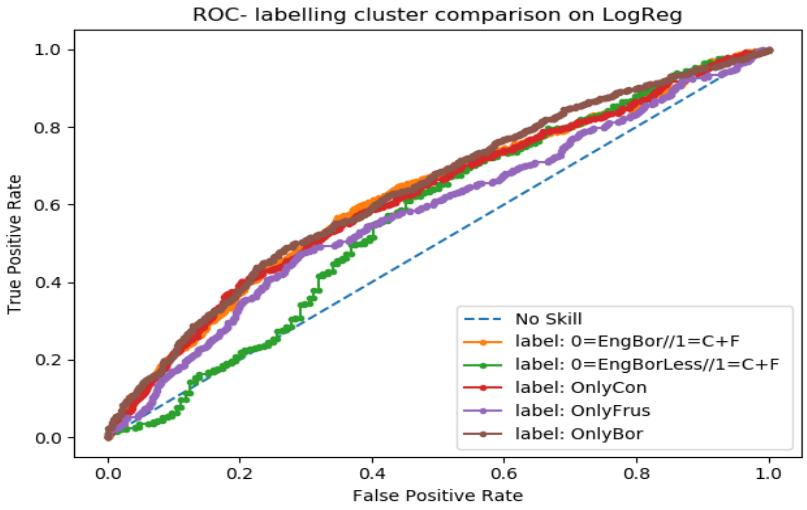
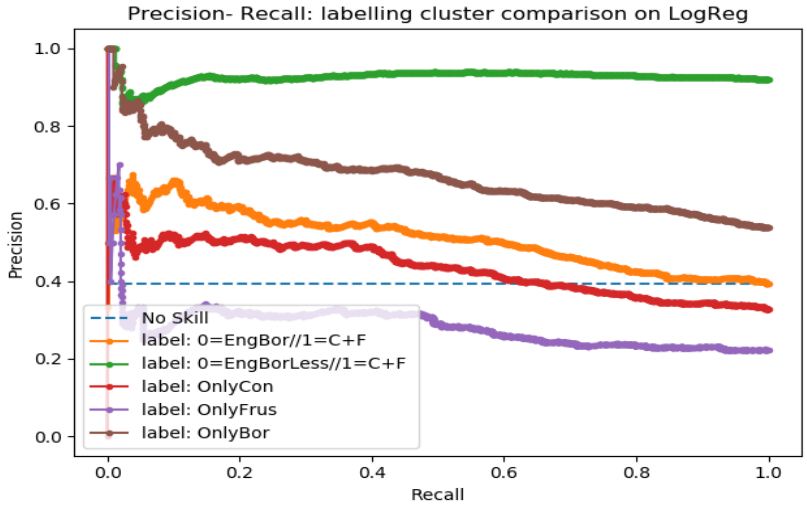


Figure 10: Precision-Recall-curve of different labelling clusters with logistic regression classifier.



Einwilligungserklärung zur Erhebung und Verarbeitung personenbezogener Daten gemäß DSGVO

Projektname: Automatic Detection of Comprehension Problems during Online Lectures
Projektleitung: Teena Hassan (thassan@techfak.uni-bielefeld.de),
Zeynep Demir (zeynep.demir@uni-bielefeld.de)
Projektdurchführung: Tatjana Maskaljunas (tatjana.maskaljunas@uni-bielefeld.de)

Datenverarbeitung

Für die Studie erfolgt die Verarbeitung folgender personenbezogener Daten:

- **Alter**
- **Geschlecht**
- **Studium/ Beruf**
- **Analysierte Gesichtsdaten**
- **Videomaterial**

Die oben genannten Daten werden zum Zweck der Analyse eines intelligenten Tutor-Systems erhoben. Lediglich die ersten vier Informationspunkte werden anonymisiert und für Dritte unzugänglich gespeichert, es erfolgt keine Weitergabe. Der Zugriff auf die Daten ist nur der oben erwähnten Projektleitung und Projektdurchführung gestattet. Die Speicherung der Daten erfolgt auf einem verschlüsselten Datenträger und die Daten werden fristgerecht nach Ablauf von **sechs Monaten** gelöscht.
Die Ergebnisse der Studie werden in einem Bericht der Projektleitung zur Verfügung gestellt und gegebenenfalls zur Veröffentlichung freigegeben. Der Bericht enthält keinerlei Video- oder Bildmaterial der StudienteilnehmerInnen und die Personendaten werden in einer anonymisierter Form erwähnt, sodass die betroffenen Personen nicht identifiziert werden können.

Widerufsrecht

Die Teilnahme an der Studie ist freiwillig. Sie haben zu jeder Zeit die Möglichkeit das Experiment abzubrechen und Ihre Einwilligung zurückzuziehen, ohne dass Ihnen dadurch irgendwelche Nachteile entstehen.
Sie haben die Möglichkeit jederzeit Ihre Einwilligung als auch Ihre Angaben zu korrigieren, diese auf Ihren Wunsch löschen zu lassen, die Nutzung der Daten zu beschränken und der gesamten Einverständniserklärung zu widersprechen. Sie haben ebenfalls das Recht sich bei rechtswidriger Nutzung ihrer Daten bei den Datenschutzbeauftragten der Universität Bielefeld zu beschweren. Ebenso haben sie jederzeit das Recht Auskunft zu erhalten welche Daten gespeichert werden, über die Art und Weise wie Ihre Daten verarbeitet und gespeichert werden und wer Ihre Daten verarbeitet.

Ich bin damit einverstanden, im Rahmen des genannten Forschungsprojektes meine oben erwähnten Personendaten zur Verfügung zu stellen.

StudienteilnehmerIn:	**Studiendurchführung:**
Ort, Datum:	**Ort, Datum:**
Unterschrift:	**Unterschrift:**

Einwilligungserklärung zur Teilnahme an der Studie

Projektname: Automatic Detection of Comprehension Problems during Online Lectures
Projektleitung: Teena Hassan (thassan@techfak.uni-bielefeld.de),
Zeynep Demir (zeynep.demir@uni-bielefeld.de)
Projektdurchführung: Tatjana Maskaljunas (tatjana.maskaljunas@uni-bielefeld.de)

Studienbeschreibung

Die Studie befasst sich mit der Fragestellung, ob ein intelligentes Tutor-System in der Lage ist Verstehensprobleme von Studierenden während einer Online-Vorlesung zu erkennen und entsprechend Hilfe anzubieten. Um dies zu realisieren werden Gesichtsdaten (Mimik, Augenbewegungen und Kopfhaltung) über die Kamera lediglich verarbeitet und dazu genutzt Rückschlüsse auf den Verstehensprozess zu ziehen.

Zunächst wird ein Erklärvideo präsentiert, welches nur angeschaut und verstanden werden soll. Im Anschluss wird eine Frage zu der vorgestellten Thematik gestellt. Während der gesamten Zeit werden die oben genannten Daten analysiert und basierend auf diesen gegebenenfalls eine Hilfestellung angeboten. Mit dem Beantworten der Frage ist die Studie beendet.

Einwilligungserklärung

Hiermit erkläre ich, dass ich über die Ziele der Studie und ihren Ablauf ausführlich und verständlich aufgeklärt worden bin. Ich hatte genügend Zeit, meine Entscheidung zur Teilnahme an der Studie zu überdenken und frei zu treffen. Ich habe eine Kopie der Einwilligungserklärung und Datenschutzerklärung ausgehändigt bekommen.

Mir ist bekannt, dass ich jederzeit und ohne Angaben von Gründen meine Einwilligung zur Teilnahme an der Studie mündlich zurückziehen kann, ohne dass mir daraus irgendwelche Nachteile entstehen.

Hiermit erkläre ich mich bereit, an der oben genannten Studie freiwillig teilzunehmen.

StudienteilnehmerIn:	**Studiendurchführung:**
Ort, Datum:	**Ort, Datum:**
Unterschrift:	**Unterschrift:**

Set	Test	Train	Validation	Train + Validation
original data	1784 [1084:700]	5412 [3297:2115]	1713 [966:747]	7125 [4263:2862]
capped data	1784 [1084:700]	5412 [3297:2115]	1713 [966:747]	7125 [4263:2862]
dropped data	1308 [824:484]	3943 [2393:1550]	1249 [713:536]	[3106:2086]

Table 5: Data.