Capturing Human Perspectives in NLP: Questionnaires, Annotations, and Biases*

Wiktoria Mieleszczenko-Kowszewicz[1,*,†], Kamil Kanclerz[1], Julita Bielaniewicz[1,†], Marcin Oleksy[1,†], Marcin Gruza[1,†], Stanisław Woźniak[1], Ewa Dziecioł[1], Przemysław Kazienko[1] and Jan Kocoń[1]

[1]Department of Artificial Intelligence, Wrocław University of Science and Technology

Abstract

This article compiles research on the extraction of human characteristics using three different methods: questionnaires, annota- tions, and biases. We have performed an analysis of how personalized perception of texts is affected by individual human profile and bias. To acquire comprehensive knowledge about individual user prefer- ences, we have gathered 40 users who annotated 1000 texts in 26 subjective tasks grouped into three categories: positive affect, neg- ative affect, and rational affect. The results revealed that categories of annotation were correlated with psychological dimensions, e.g., agreeableness and conscientiousness, which are traits related to pos- itive affect dimension biases. We have observed the presence of two clearly defined categories among annotators when it comes to the aspect of humor: those who confidently share their perspectives on what they find funny and those who tend to rate humor levels within a narrow range. Moreover, we analyzed intra-annotator agreement to show that people tend to change their ratings over time. Our results show that the higher level of the ranking correlation between anno- tations and agreement calculated using binarized annotations com- pared to the absolute agreement calculated using full annotations im- plies that the 10-point annotation scale might be a significant factor in annotator disagreement.

Keywords

natural language processing, personalization, subjectivity, annotator bias, annotator representation, data acquisition,

1. Introduction

Resolving natural language processing (NLP) tasks, such as detecting offensiveness, humor recognition, or emotion recognition, requires the work of annotators labeling large datasets used in training models in machine learning algorithms. Although people vary between themselves on a daily basis, the final evaluation of annotated instances is a decision of the majority of the annotator called *the gold standard*. The assumption underlying this process is that most people will perceive texts similarly [1]. Annotations not aligned with the majority vote are not included in the final model. As a result, much information about humans is not used. Moreover, annotators' personalities are flattened and generalized, affecting the model's accuracy. Despite existing research [2, 3], there is still a certain lack of exploration in measurement of the way how individual characteristics of the text's audience influence the perception of it.

This article aims to answer the following research questions:

1. What is the impact of annotators' individual characteristics on their text perception?
2. How does the evaluation of texts change over time and what are the crucial factors of such an intra-annotator change of the user?
3. What are the main differences between methods for capturing human perspectives?
4. What is the impact of annotator sense of humor on the funny content perceived by themselves and other people?
5. What are the ranking dependencies of annotations and absolute agreement between annotators?

2. Related Work

The research from recent years has shown that people strongly vary in their perception of text depending on the characteristics they possess. This includes features such as cognitive skills [4], personality traits [5], or even the emotions they have experienced [6]. This noticeable diversity between people is reflected in the multiple perspectives presented in the annotations. The work of Basile et al. [7] states that the perspectivist approach should be taken into account when determining the golden standard. What it implies is the need to tailor the standard to each person individually, understanding that the said ground truth is subjective. As the differences in user reception of the same text inevitably

2nd Workshop on Perspectivist Approaches to NLP

*Corresponding author.

†These authors contributed equally.

wiktoria.mieleszczenko-kowszewicz@pwr.edu.pl (W. Mieleszczenko-Kowszewicz)

become apparent, it is crucial to examine it using appropriate measures [8]. A stability of user's annotations is an interesting take, however we have decided to focus on the deviating from the majority. For this reason, we have utilized measures such as Personal Emotional Bias [9] and Human Bias [10]. The first metric calculates the degree of user differentiation from the average emotional perception of a given text, while the second metric compares the bias of an annotator and its similarity to the majority of users. As seen over the years, applying these measures when performing experiments in natural language processing tasks [11, 12, 13, 14] confirmed the effectiveness and a strong improvement in understanding the individuality of a user. Furthermore, it has been shown that compared to standard methods derived from psychology, NLP models are even better at identifying the Big Five personality traits [15]. With that in mind, we have decided to perform an assessment of results from a collection of different questionnaires, as well as investigate the annotations of users.

3. Capturing Human Perspectives

3.1. Text Selection Procedure

To acquire comprehensive knowledge about individual user preferences, our annotation process consisted of three major steps: (1) annotation of the large collection of texts done by a small group of annotators (6 people), (2) measuring the controversy of the annotated texts with three methods, and (3) selection of texts for annotation involving a large group of users (40 people). In the first step, a small group of experienced annotators annotated a large collection of comments in Polish. They were acquired from various Internet forums regarding news, sport, and lifestyle topics. Then, we measured the controversy [12] of texts in 3 variants: (1) average controversy for all dimensions, (2) average controversy of the top five most controversial dimensions for the specific text, and (3) highest controversy value of all dimensions for a certain text. Finally, we separately selected $\frac{1}{3}$ of the texts for annotation with each variant of the controversy. Furthermore, the texts selected by a specific variant consisted of $\frac{2}{3}$ of the texts with the highest controversy and $\frac{1}{3}$ of the texts with the lowest controversy measured by a specific variant. In this way, the final dataset obtained in step (3) comprised texts with diverse controversy, which enabled the extraction of various user perspectives.

3.2. Dataset

Forty annotators participated in the study, with 77.5 % of them being women and 22.5% being men. Their age ranged from 19 to 56 years (μ = 39.9, σ = 10.1).

Table 1
Annotation dimensions categorized depending on the affect and rational nature.

Positive affect	Negative affect	Rational (no affect)
(1) calm	(8) anger	(13) agreement
(2) compassion	(9) disgust	(14) embarrassing
(3) delight	(10) fear	(15) funny to me
(4) inspiration	(11) negative	(16) funny to someone
(5) joy	(12) sadness	(17) incomprehensible
(6) positive		(18) interesting
(7) surprise		(19) ironic
		(20) offensive to me
		(21) offensive to someone
		(22) political
		(23) sympathy
		(24) trust
		(25) understandable
		(26) vulgar

The dataset we used is one of the iterations of the Doccano 1.0 project, which aims to capture subjective impressions elicited by textual content. The number of annotated texts was 1000. Each of them is no longer than 132 words (μ = 24.5, σ = 16.2). On average, each person annotated around 790 texts and each text was annotated by around 32 annotators. In its entirety, it comes out a little under 31,700 annotations. Each annotation consists of 26 independent dimensions (see Tab 1: For each dimension, the annotator chose a value from 0 to 10, where 0 means that the annotator did not react and 10 means that the reaction was strong. No decision is acceptable, indicating that the person does not know what value to give. Labels with a value of zero occur on average 62% with 22% standard deviation in each dimension. Meanwhile, empty labels occur on average 4% with 8% standard deviation. The distributions of the remaining values, which provide us with information about the actual reactions of the annotators, are shown in Fig. 1.

The dimensions are divided into three groups: positive affect, negative affect, and rational (no affect). This approach is inspired by multiple works [16, 17, 18].

3.3. Measuring Annotator Profile: Questionnaires

Big Five personality traits (Mini-IPIP) [19] is a 20 item questionnaire that measures the factors of the Big Five personality model: *extraversion*, *agreeableness*, *conscientiousness*, *neuroticism*, and *intellect/imagination*. Each dimension is measured by four questions, where answers are given on a 5-point scale: 1 = very inaccurate to 5 = very accurate. *Agreeableness* is considered a social trait that aims to maintain positive relationships with others. People who score high on this trait tend to choose the interpretation of the situation as less controversial

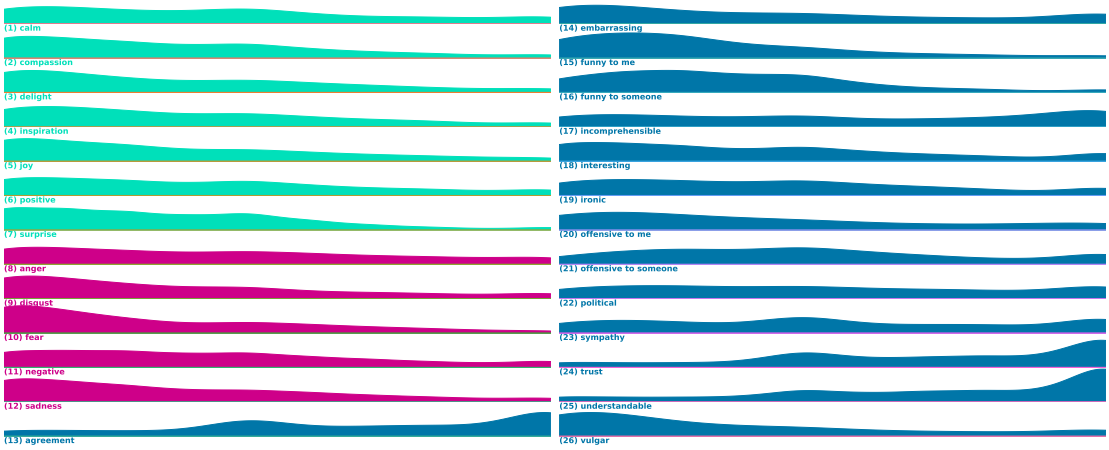

Figure 1: Distribution of non-zero values for dimensions in Tab. 1: positive affect, negative affect, and rational (no affect).

and choose the more constructive form of conflict resolution [20]. *Extraversion* is a trait that describes people who are active and social, it is also widely known for its association with positive affect. *Conscientiousness* is a personality characteristic that describes the tendency to be organized, prepared, hard working, and maintaining a high quality of work [21, 22]. *Neuroticism* refers to the tendency of people to experience negative emotions such as anxiety, worry, fear, and sadness [23]. *Intellect* is a trait that describes the willingness to seek new experiences, investigate new ideas, experience new tastes, and visit new places [24].

Humor Styles Questionnaire (HSQ) [25] is a 32 element questionnaire that evaluates four styles of humor applied by a person: (1) *self-enhancing*, (2) *affiliative*, (3) *aggressive*, and (4) *self-defeating*. The two positive values indicate (1) the empowerment of self through the use of humor and (2) the willingness to bond with others (mostly the recipients of the texts). The remaining negative values refer to (3) inflicting a verbal attack on other people, as well as (4) themselves through the use of deprecating humor. The values of each of the styles are calculated through the use of answers to 32 questions regarding the sense of humor of an individual, which includes 8 questions per individual style of humor. The scale of answers consists of 7 possible answers from 1 = totally disagree to 5 = totally agree.

The regulating emotion systems in everyday life (RESS-EMA) scale [26] evaluates how people regulate their emotions in daily life. The questionnaire consists of 12 items measuring 6 emotion regulation strategies (2 items per subscale). Each item was rated on scales from 0 = totally disagree to 100 = totally agree, and the respondents ticked off which emotion management strategies they had used in the past month. The subscales are: *relaxation* (dampening of autonomic arousal), *engagement* (active expression of emotions), *rumination* (sustained attention), *reappraisal* (cognitive reframing), *distraction* (diverting attention) and *suppression* (inhibition of emotional expression).

The Physical Health Questionnaire PHQ [27] is a 14-item questionnaire that evaluates four dimensions of somatic health (*sleep disturbances*, *headaches*, *gastrointestinal problems* and *respiratory infections*). Items were rated on a 7-point frequency scale with seven possible answers.

Patient Health Questionnaire-9 PHQ-9 [28] is a questionnaire consisting of 9 questions about the symptoms of *depression*, which the user rates on a scale of 0 to 3.

Depression is one of the most common mental disorders. The core questions of the PHQ-9 address the symptoms of depression included in the DSM-IV diagnostic criteria: the higher the score, the more severe the depression.

In **PHQ** and **PHQ-9** questionnaires, the lowest scores correspond to the absence of symptoms, while the higher scores proportionally represent their more frequent occurrence.

Alexithymia measured with the PAQ questionnaire [29] containing 7 questions on the 7-point Likert scale [30] ranges from 1 = strongly disagree to 7 = strongly agree. It is a trait that impedes identifying own feelings, describing them, and limits externally oriented thinking style, manifesting in unintentionally ignoring others' emotions.

Perceived Stress Scale [31] measures stress with 10 items on 5-point scale with answers from 0 = never to 4

= very often.

Scale of Positive and Negative Experience [32] is a 12-item questionnaire that measures positive and negative feelings with two subscales (6 items each). Possible answers range from 1 = very rarely or never to 5 = very often or always.

The Satisfaction with Life Scale (SWLS) [33] 7 items questionnaire measures global life satisfaction. Users can answer each question on a 7-point Likert scale ranging from 1 = strongly disagree to 7 = strongly agree.

3.4. Human Bias

We used the Human Bias $\mathrm{HB}(u, d)$ [14] measure to capture the diversity between the preferences of the user and the others. Its value for a user u within dimension d is a Z-score-based measure that describes the degree of diversity of user u's annotations $v_{d,t,u}$ of all texts $t \in T_u$ relative to the mean $\mu_{d,t}$ and standard deviation $\sigma_{d,t}$ of annotations provided by all users in dimension d, as follows:

$$\mathrm{HB}(u, d) = \frac{\sum_{t \in T_u} \frac{v_{d,t,u} - \mu_{d,t}}{\sigma_{d,t}}}{|T_u|} \quad (1)$$

3.5. Back Saturation

For the purposes of the study, we introduced a measure called Back Saturation (BS). It could be calculated for each text (T) within a particular dimension as follows:

$$BS_N = N_{-1} * 3 + N_{-2} * 2 + N_{-3} + N_{-4} + N_{-5} \quad (2)$$

where N is the rating for the negative dimension. -1 refers to the one text back, -2 to the two texts back, etc., for example, if subsequent texts received the following negativity ratings: T1 - 3, T2 - 5, T3 -3, T4 - 2, T5 - 7, then the BS_N for text T6 is:

$$BS_{N_T 6} = 7 * 3 + 2 * 2 + 3 + 5 + 3 = 36 \quad (3)$$

3.6. Intra-Annotator Agreement

Inspired by the recent works [8] we randomly selected 3 annotators for a very detailed analysis. Its purpose was not only to examine the consistency of the annotations, but also to try to determine the influence of various factors on the change of their decisions. The annotation process was planned in such a way that some texts appeared at least twice (hereinafter 'duplicates'). It was then possible to calculate the consistency of the annotations of these texts made by a single annotator (hereinafter 'intra-annotator agreement' or IntraAA). For some purposes we have also introduced soft IntraAA, where annotations that differ by only one point (on a scale of 1-10) are also considered as consistent.

4. Analytical Results

4.1. Annotator Profile

Regarding personality traits, the participants displayed moderate levels of agreeableness, conscientiousness, and intellect, while exhibiting slightly lower levels of extraversion and neuroticism. In terms of humor styles, affiliative and self-enhancing humor was commonly observed, whereas aggressive and self-defeating humor styles were less prevalent. Participants exhibited relatively high engagement and reappraisal in stress regulation, indicating a tendency to express emotions and actively reinterpret them. Furthermore, they have also reported relatively low levels of alexithymia. In addition, participants reported a moderate level of perceived stress with a tendency to experience positive rather than negative affect. On average, participants reported a moderate level of life satisfaction. The detailed characteristics of the annotators are presented in Appendix A.

4.2. Bias and Human Characteristics

Personality described in Appendix B.1 shows the correlations between the Big Five and the annotations. The results reveal that agreeableness and conscientiousness are traits that are strongly related with positive affect biases. Slightly weaker tendency is observed for negative affect biases. Moreover, these two traits are also moderately correlated with each other, which strengthens the above observation.

Styles of humor and subjectivity. Despite the fact that every human understands the concept of humor, each person has their own, distinct sense of it. We can analyze the similarity between each person, aggregate the annotation scores into groups, and eventually find the humor scores of the majority of annotators, but there is a very low chance of encountering people with identical set of scores related to humor. Even so, the same scores in this particular research would not imply that the annotators with equally same humor annotations possess the exact same sense of humor. This indicates the fact that humor is a hugely subjective task, and with this in mind, we need to take into account the perspective of the individual user when assessing their results. As humor in natural language processing itself is a vastly personalized task, identifying and categorizing texts with different types of humor may shed some light on the details of a person's sense of humor. The categorization derived from the Humor Styles Questionnaire in Sec. 3.3 provides a set of humor types that are widely used in the field of humor research, not only in the scope of natural language processing, but also in psychology [34, 35]. When acquired, the four available measures of different types of humor indicate the intensity of experiencing

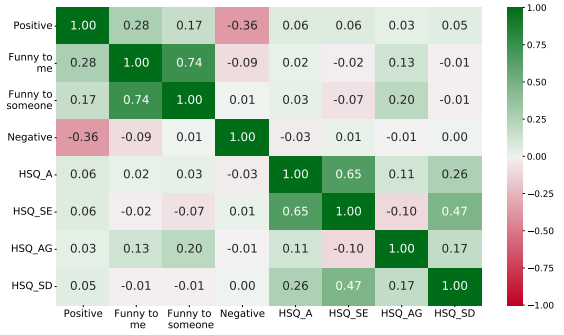

Figure 2: Correlation of annotator-based biases and answers in humor dimension. As seen through the low correlation values, the Humor Style Questionnaire seems to dismiss the perspective of user themselves and focus on the view of themselves in other people.

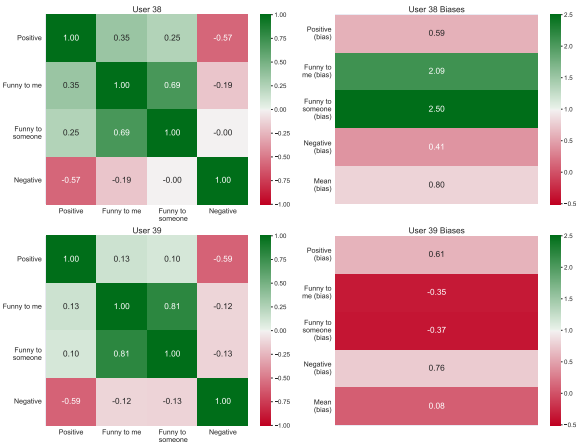

Figure 3: Correlation of biases and answers in the humor dimensions. The first row presents User number 38, who freely expresses their views on funniness, as where the User 39 from the second row keeps his emotions at bay, often not finding texts funny. This phenomenon is noticed through the correlation values in the figure.

humor, but what is interesting is that we can see in Fig. 2 that these values actually focus on the external perspective of funniness of an individual. It is clearly visible when evaluating the correlation values between the humor style parameters and the dimensions *funny to me* and *funny to someone*. as presented in Fig. 3. Other than the mentioned results, the HSQ metrics seem to be separated from the standard funniness values, as the correlation is much lower than when analyzed between other HSQ values. As for the individualism of a user, the subjective matter of experiencing humor is based on the emotionality of a user. We have noticed that there are two distinct groups of annotators in regard to the humor dimension, people who feel free to express their views of funniness, and individuals who hardly exceed small values in both funniness and unfunniness. As shown in Fig. 3, people from the expressive group, such as User 38, have a relatively high correlation when talking about the content being *funny to others* or themselves, as where more reserved people, similar to User 39, tend to be mild in their expression of emotions and feelings. This observation extends the area of subjectivity in humor in NLP and emphasizes that not only the experience is analyzed through personalization, but also the expression must be noticed and thoroughly examined. Detailed correlation between humor and annotation dimensions is presented in Appendix B.2.

Emotion regulation and subjectivity The relationships between regulation of emotions and subjectivity are described in the appendix B.3. The use of distraction as a strategy exhibits the most positive relationship with the positive affect dimension and the selected rational biases. On the contrary, the relaxation strategy shows an inverse relationship with negative rational biases.

Health and subjectivity is described in Appendix B.4. Depression and gastrointestinal problems are the health characteristics that are more correlated with the negative affect dimension. A similar relationship exists between vulgar and embarrassing bias. Also, compassion is positively related to health problems. There is a general tendency for people who report health problems to perceive text as less understandable.

Bias and Stress with Emotions are presented in Appendix B.5. Stress is related to positive and slightly weak to negative affect biases. On the other hand, there is a negative relationship between experiencing positive affect and rational biases. Negative affect is related to negative affect and rational dimensions. Satisfaction with life is weakly negatively related to rational dimension biases.

4.3. Intra-Annotator Agreement over Time

The sample results (calculated for one annotator)[1] are shown in Fig. 4. Interestingly, IntraAA only in few cases reaches a level that could be considered very good, or even satisfactory. The situation is even worse if we exclude the cases of the agreement for null marks, especially when they account for a large percentage of decisions (e.g. for the presented user the score ranges from 0.08 to 0.54, and the average is 0.21). However, the use of the soft IntraAA, which also considers as congruent answers those that differ only by one point, shows that the differences between the annotations are most often not

[1]For the complete results for all 3 annotators see the Appendix C

large - the IntraAA increases significantly (55% on average for the analyzed users). This shows that the analyzed annotators were characterized by relatively high stability. Smaller differences between strict and soft IntraAA would show the dimensions for which annotators are particularly stable. Such dimensions include *joy*, *inspiration*, *embarrassing*, *vulgar* or *offensive to me*.

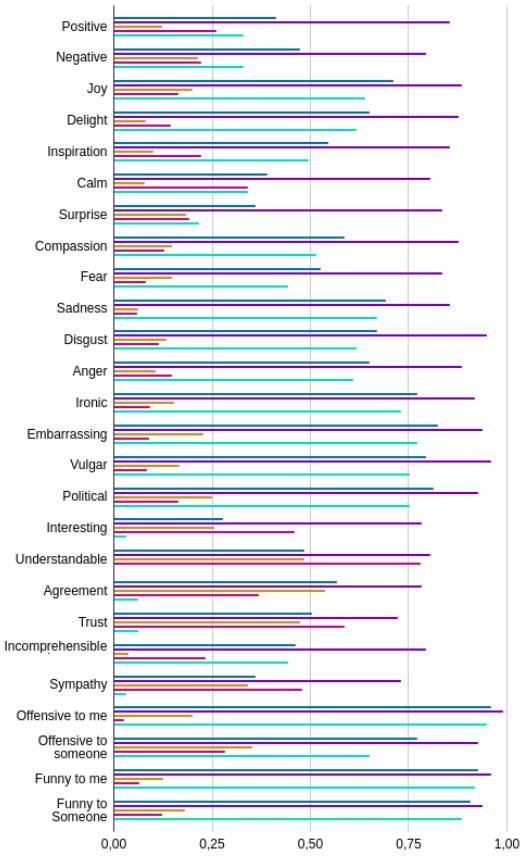

Figure 4: Intra-Annotator Agreement calculated for a selected user (id 0). Blue bars represent strict IntraAA, violet bars represent soft IntraAA, yellow bars represent the IntraAA calculated with the exclusion of the consistent annotations for zero labels, pink bars are for the average rating for each dimension, and the turquoise bars represent the proportion of zero markings.

We believe that a number of factors can affect the change in rating. The basis for the more detailed analysis was the labels within the *negative* dimension, primarily because this is the dimension for which relatively most labels other than "zero" appear and because it has relatively low concordance scores. Among other things, the analysis looked at the impact of time. It turned out that the tendency to change the decision increased when the text to be annotated was repeated on a different day (some duplicates appeared on the same day). Interestingly, for decisions made on two different days, the proportion of changes from a more to a less negative label increased (at the expense of cases of maintaining the assessment; see Fig. 5).

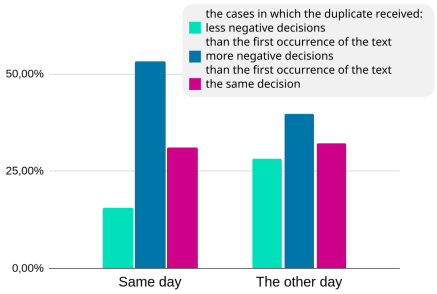

Figure 5: Changing the decision over time for the *Negative* annotation dimension.

We also investigated this phenomenon by trying to determine the impact of the negativity of previously annotated texts. For the purposes of the study, we used a measure called Back Saturation (BS - see Section 3.5). After assigning the appropriate BS_N value to each text, we compared respectively the BS_N for each text as it appears for the first time and for the second time. The results were combined with changes in the annotator's decision (see Fig. 6). As it turns out, the analyzed annotators changed their decisions without a clear effect of back saturation. However, we observe an imbalance in the proportion in the case where the evaluation of a text changes to a more negative text by one point. Indeed, we note relatively more cases in which such a decision change is associated with the occurrence of a duplicate after more negative texts.

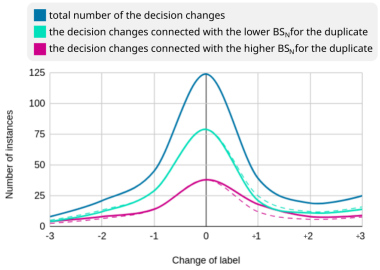

Figure 6: The correlation between BS_N and changes in the decision.

4.4. Relation between Annotations, Questionnaires and Biases

To gather holistic knowledge about the user, we decided to include text annotations and questionnaires in the data

Table 2
Three methods of human data acquisition.

Category	Questionnaires	Annotations	User biases
text dependency	none	large	medium
effort	medium	large	small
time	medium	large	small
cost	medium	large	small

collection process. Then, we used the acquired annotations to calculate the biases that describe the peculiarity of user preferences according to others. Each of the human data acquisition methods are described in Tab. 2. To measure the similarity of knowledge obtained by each of these methods, we used the Pearson correlation coefficient [36]. The results are presented in Fig. 7. The higher correlation values were observed between text annotations and user biases. On the other hand, lower correlation values appeared between questionnaire answers and user biases. The relation between questionnaires and text annotation is described by the least significant correlation values.

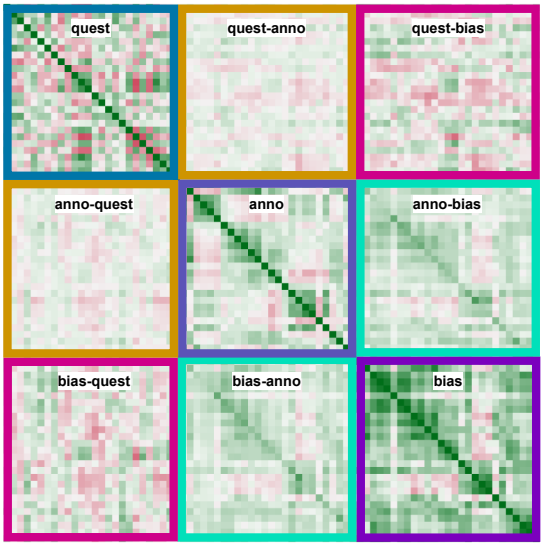

Figure 7: Correlation between annotation dimensions (anno), questionnaires (quest), and user biases (bias). Regions outlined with the same color are each other transposes. Green cells mark the positive correlation, white cells mark no correlation and the red cells mark the negative correlation values. Cell color intensity describes correlation values: more green – higher positive correlation, more red – lower negative correlation.

4.5. Inter-Annotator Agreement

The averaged agreement measures between the annotators in the dataset are provided in Tab. 3. The multivalued 10-point scale for each dimension makes it difficult to achieve exact agreement between annotators. Therefore, to better understand the phenomenon, we used three different agreement metrics:

1. **Cohen's kappa** on raw annotations.
2. **Cohen's kappa on binarized annotations**, where all nonzero annotations (1-10) were converted to ones (1).
3. **Kendall Tau rank correlation coefficient** that measures the ranking agreement between annotators.

In case of Kendall rank correlation metrics, all empty annotations were removed from calculations, as they cannot be ordered. As expected, the average Kappa agreement scores in most dimensions are very low, with a minimum for *surprise* (0.025) and maximum for *political* (0.267). In the case of binarized annotations, the Kappa agreement increases significantly (between 0.052 for *surprise* and 0.513 for *political*). The Tau coefficient ranges between 0.081 for *surprise* and 0.589 for *political*. The results also reveal a positive correlation between the percentage of zero annotations and annotators agreement for given dimension (0.485 Pearson correlation coefficient for the mean kappa and 0.396 for mean kappa binarized). We also checked the correlation between the mean Tau coefficient and the absolute differences in the biases of the annotators. Annotators with high bias are more likely to rate texts above average, and annotators with low bias are more likely to rate texts below average. Therefore, the difference of biases on given dimension can be interpreted as the distance between the annotators' sensitivity on this dimension. As Tab. 3 shows, these correlations are mostly negative but very weak. This means that there is no clear relationship between the annotator ranking agreement and the difference in their sensitivity.

5. Discussion

The results of our research revealed that human characteristics have an influence on biases. Higher levels of agreeableness and conscientiousness are associated with increased differentiation in *positive dimension biases*. Individuals with lower levels of neuroticism tend to exhibit stronger biases toward positive affect dimensions, and individuals with higher levels of intellect may exhibit weaker biases toward positive affect dimensions. Positive affect dimension biases are represented by individuals who primarily rely on distraction strategy. Individuals who engage in aggressive and self-defeating humor tend to be more controversial in their biases toward positive affect dimensions, suggesting that humor styles may influence the perception and interpretation of positive emotional perception of the text. Individuals

Table 3
Agreement averaged over all pairs of annotators, on all dimensions. Agreement was calculated with two metrics: Kendall rank correlation coefficient and Cohen's kappa.

Dimension	Mean kappa	Mean kappa binarized	Mean tau	Tau-kappa corr.	Tau-bias diff. corr.	Zero annotations
positive	0.125	0.275	0.336	0.718	-0.127	0.583
negative	0.117	0.347	0.365	0.562	-0.084	0.421
joy	0.130	0.258	0.296	0.816	-0.108	0.763
delight	0.147	0.281	0.321	0.877	-0.116	0.826
inspiration	0.151	0.300	0.339	0.881	-0.258	0.753
calm	0.097	0.193	0.253	0.744	-0.234	0.711
surprise	0.025	0.052	0.081	0.718	-0.039	0.622
compassion	0.092	0.195	0.242	0.775	-0.325	0.688
fear	0.103	0.205	0.238	0.810	-0.072	0.769
sadness	0.103	0.235	0.267	0.775	-0.113	0.643
repulsion	0.127	0.268	0.295	0.790	-0.162	0.747
anger	0.134	0.325	0.335	0.740	-0.157	0.590
ironic	0.135	0.275	0.342	0.785	-0.341	0.633
embarrassing	0.140	0.270	0.319	0.852	-0.065	0.758
vulgar	0.244	0.451	0.501	0.823	-0.319	0.853
political	0.267	0.513	0.589	0.755	-0.256	0.760
interesting	0.045	0.116	0.179	0.654	-0.125	0.402
understandable	0.029	0.056	0.181	0.492	-0.272	0.111
incomprehensible	0.051	0.101	0.167	0.683	-0.150	0.482
offensive to me	0.070	0.126	0.147	0.929	0.032	0.960
offensive to someone	0.147	0.328	0.363	0.793	-0.261	0.680
funny to me	0.086	0.156	0.181	0.898	-0.149	0.908
funny to someone	0.100	0.193	0.230	0.887	-0.199	0.848
trust	0.070	0.157	0.249	0.587	-0.203	0.192
sympathy	0.056	0.129	0.277	0.581	-0.214	0.296
agreement	0.088	0.165	0.281	0.571	-0.213	0.223

who are in a positive emotional state are more likely to perceive and interpret stimuli in a positive light. Individuals who have higher levels of life satisfaction may have a generally positive outlook, influencing their perception and interpretation of stimuli as more positive. Generally, according to questionnaire data, there is a tendency that positive affect dimensions are affected by the level of health (both mental and physical). Interestingly, people with health problems evaluate text as more arousing compassion.

Individuals who do not use relaxation strategies as a coping mechanism for stress tend to exhibit a *negative affect dimension bias*. This suggests that the absence of relaxation techniques may contribute to a tendency to perceive and interpret stimuli in a negative light when experiencing stress. Individuals who employ affiliative humor are less likely to present biases toward negative affect dimensions. There is a positive relationship between agreeableness and differentiation in negative dimension biases, slightly weaker compared to positive dimension biases. This implies that individuals with higher levels of agreeableness may display more nuanced biases when it comes to perceiving negative affect. Higher scores in alexithymia are associated with a greater propensity to negative bias. This suggests that individuals who struggle with identifying and expressing their own emotions may be more inclined toward negative biases in their perception and interpretation of stimuli. When individuals experience negative emotions, they are more susceptible to perceiving text through a negative dimension bias. This implies that the emotional state of negativity can influence how individuals interpret and evaluate stimuli, leading to a bias towards negative affect dimensions. Individuals who report lower levels of life satisfaction tend to mark text as more negatively biased. This finding suggests that lower life satisfaction may contribute to perceiving and evaluating stimuli in a negative light, influencing negative dimension biases in the interpretation of the text. People with health problems are more prone to negative dimension biases. Surprisingly, affective biases are less noticeable when people experience stress. *Vulgar* and *embarrassing* bias co-occur with each other. People who score higher in neuroticism, experiencing stress, feeling negative emotions, and less satisfied with life are more prone to perceive texts as more controversial in those two biases. Depression and general health problems can reinforce these biases, as well as rumination and distraction as emotion regulation strategies. An inverse relationship with positive emotions confirms the tendency to perceive text as passing less controversial while experiencing similar emotions. A similar tendency is noticed for people who score higher in intellect and use relaxation and engagement as strategies of emotion regulation. Individuals who are more likely to view a text as *offensive* or *funny* tend to experience higher levels of stress, negative emotions, and difficulty in identifying and understanding their own emotions. Additionally, the presence of positive affect appears to have a mitigating effect on this tendency, indicating that higher levels of positive emotions are associated with a reduced likelihood of perceiving the text as offensive or funny. There is a difference between personality traits that have an impact on the *offensive to me* and *offensive to someone* bias. Individuals who score higher in agreeableness and conscientiousness have a tendency to perceive the text as more *offensive* to them, surprisingly the tendency is inverse for an offensive to someone (only for agreeableness). In other words, individuals high in agreeableness and offensiveness may be more sensitive to personal criticism or offensive remarks directed toward them, but

they may be less sensitive or more understanding when it comes to offensive language or content directed toward others. There is also a positive relationship between perceiving text as *offensive to someone* and *funny* (to me or someone) with the rumination and suppression strategy. Interestingly, no significant relationship was observed between the rumination strategy and the perception of text as *offensive* to oneself, suggesting that this particular strategy may not significantly influence one's sensitivity to personal offense. The use of distraction as a coping mechanism has an impact on perceiving content as offensive and finding humor in it. The inverse relationship is observed for conscientiousness. Individuals with higher levels of conscientiousness may be more sensitive to potential threats or negative implications in communication, leading them to perceive text as offensive to them more frequently. Individuals with higher levels of intellect are less likely to interpret text as personally *offensive*. In other words, intellectual individuals tend to be more objective and less sensitive to potentially *offensive* content directed at themselves. *Political bias* is higher for people who score higher in intellect. The same tendency is for neuroticism. Individuals who are more agreeable are likely to be a more open-minded and tolerant approach when it comes to political beliefs, leading to lower levels of political bias. Also, health problems can influence the perception of text as understandable. However, to generalize such conclusions, we should conduct more complex studies that consider the use of more specialized equipment. The fact that the ranking agreement (Tau coefficient) and agreement calculated on binarized annotations (Kappa binarized) are significantly higher than the agreement calculated on raw annotations suggests that the 10 point annotation scale may be problematic for annotators. They generally agreed on the presence of a given dimension in the text, but differed in determining its exact intensity. Nevertheless, the values of the Tau coefficient are high for most of the tasks, which means that the annotators generally agreed on the ranking of the dimension intensity of texts.

Higher correlation values between text annotations and user biases compared to their relationship with questionnaires may be related to the text dependency of those methods. On the other hand, more significant positive and negative correlations between questionnaires and biases compared to correlations between questionnaires and annotations may be caused by the aggregative nature of biases. They aim to distill user annotations to emphasize the main differences between user preferences compared to others. Furthermore, the highest number of negative correlation values was observed between questionnaires and biases. This outlines the different types of text-agnostic knowledge about the user that can be obtained with this method in comparison to annotations and biases. Therefore, the distinct nature of those methods implies the necessity to include all of them in the data acquisition process in order to capture the most relevant representations of various human perspectives.

The analysis of the stability of the ratings showed several important issues. The difference in the evaluation of duplicates made on a different day than the annotation of the first occurrence of the text may indicate a gradual resilience to the content presented since users rather lowered the score for negativity than upheld their judgment. The introduction of a new measure to determine the negativity of the context in the form of preceding texts (BS_N) revealed that there is an impact of the negativity of texts previously rated by the annotator – if the context for the duplicate is more negative than for the first occurrence of the text (BS_N is higher), the annotators tend to assign a more negative rating to the duplicate than they did for the first appearance.

6. Conclusions and Future Work

Our results demonstrated that people vary between themselves in terms of psychological characteristics, which was also reflected in the diversified annotation results. Relationships between questionnaire results and biases lead to several conclusions. First, there is a common tendency that specific psychological characteristics are related to similar dimensions inside the group. e.g., agreeableness with positive affect. It is a question of future research to investigate why certain dimensions (e.g., calm with agreeableness) did not correspond to the group tendencies. Second, it is possible to evaluate the intensity of psychological characteristics based on the annotation of texts. Future studies could further explore this issue by selecting the type of text to annotate and developing population norms. The main conclusion that can be drawn is that psychological characteristics influence multiple perspectives on text perception. Our research also shows that it may be worth including information about annotator characteristics in machine learning solutions. We have shown that people tend to change their ratings over time, and in many cases, the differences in annotations (and therefore intra-annotator agreement) are very high. Undoubtedly, this depends on many factors. One of them may be the influence of previously annotated texts. We presented a study conducted by us on a selected sample of annotators. Our future work in this regard would involve increasing the scope of this work to more dimensions and a larger number of annotators. The limitations of the present studies naturally include the unbalanced gender and age group. Another limitation concerns insufficient sample size to generealize our findings. The source code used during research is publicly available[2].

[2]https://github.com/CLARIN-PL/capturing-human-perspectives/tree/main

Acknowledgments

This work was financed by (1) the National Science Centre, Poland, project no. 2021/41/B/ST6/04471; (2) Contribution to the European Research Infrastructure 'CLARIN ERIC - European Research Infrastructure Consortium: Common Language Resources and Technology Infrastructure', 2022-23 (CLARIN Q); (3) the Polish Ministry of Education and Science, CLARIN-PL; (4) the European Regional Development Fund as a part of the 2014-2020 Smart Growth Operational Programme, projects no. POIR.04.02.00-00C002/19, POIR.01.01.01-00-0288/22 and POIR.01.01.01-00-0923/20; (5) the statutory funds of the Department of Artificial Intelligence, Wroclaw University of Science and Technology; (6) the Polish Ministry of Education and Science within the programme "International Projects Co-Funded"; (7) the European Union under the Horizon Europe, grant no. 101086321 (OMINO). However, the views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor European Research Executive Agency can be held responsible for them.

References

[1] D. Hovy, S. Prabhumoye, Five sources of bias in natural language processing, Language and Linguistics Compass 15 (2021) e12432.

[2] K. Kenyon-Dean, E. Ahmed, S. Fujimoto, J. Georges-Filteau, C. Glasz, B. Kaur, A. Lalande, S. Bhanderi, R. Belfer, N. Kanagasabai, et al., Sentiment analysis: It's complicated!, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2018, pp. 1886–1895.

[3] A. M. Davani, M. Díaz, V. Prabhakaran, Dealing with disagreements: Looking beyond the majority vote in subjective annotations, Transactions of the Association for Computational Linguistics 10 (2022) 92–110.

[4] A. Tourimpampa, A. Drigas, A. Economou, P. Roussos, Perception and text comprehension. it'sa matter of perception!, International Journal of Emerging Technologies in Learning (Online) 13 (2018) 228.

[5] M. M. Nitzschner, U. K. Nagler, J. F. Rauthmann, A. Steger, M. R. Furtner, The role of personality in advertising perception: An eye tracking study, Psychologie des Alltagshandelns 8 (2015) 10–17.

[6] X. Sun, X. Zhou, Q. Wang, S. Sharples, Investigating the impact of emotions on perceiving serendipitous information encountering, Journal of the Association for Information Science and Technology 73 (2022) 3–18.

[7] V. Basile, F. Cabitza, A. Campagner, M. Fell, Toward a perspectivist turn in ground truthing for predictive computing, arXiv preprint arXiv:2109.04270 (2021).

[8] G. Abercrombie, V. Rieser, D. Hovy, Consistency is key: Disentangling label variation in natural language processing with intra-annotator agreement, arXiv preprint arXiv:2301.10684 (2023).

[9] P. Milkowski, M. Gruza, K. Kanclerz, P. Kazienko, D. Grimling, J. Kocon, Personal bias in prediction of emotions elicited by textual opinions, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop, Association for Computational Linguistics, Online, 2021, pp. 248–259. URL: https://aclanthology.org/2021.acl-srw.26. doi:10.18653/v1/2021.acl-srw.26.

[10] P. Kazienko, J. Bielaniewicz, M. Gruza, K. Kanclerz, K. Karanowski, P. Miłkowski, J. Kocoń, Human-centred neural reasoning for subjective content processing: Hate speech, emotions, and humor, Information Fusion (2023).

[11] J. Bielaniewicz, K. Kanclerz, P. Miłkowski, M. Gruza, K. Karanowski, P. Kazienko, J. Kocoń, Deepsheep: Sense of humor extraction from embeddings in the personalized context, in: 2022 IEEE International Conference on Data Mining Workshops (ICDMW), 2022, pp. 967–974. doi:10.1109/ICDMW58026.2022.00125.

[12] K. Kanclerz, A. Figas, M. Gruza, T. Kajdanowicz, J. Kocon, D. Puchalska, P. Kazienko, Controversy and conformity: from generalized to personalized aggressiveness detection, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 5915–5926. URL: https://aclanthology.org/2021.acl-long.460. doi:10.18653/v1/2021.acl-long.460.

[13] K. Kanclerz, M. Gruza, K. Karanowski, J. Bielaniewicz, P. Miłkowski, J. Kocoń, P. Kazienko, What if ground truth is subjective? personalized deep neural hate speech detection, in: Proceedings of the 1st Workshop on Perspectivist Approaches to NLP@ LREC2022, 2022, pp. 37–45.

[14] J. Kocoń, M. Gruza, J. Bielaniewicz, D. Grimling, K. Kanclerz, P. Miłkowski, P. Kazienko, Learning personal human biases and representations for subjective tasks in natural language processing, in: 2021 IEEE International Conference on Data Min-

ing (ICDM), IEEE, 2021, pp. 1168–1173.
[15] A. Cutler, D. M. Condon, Deep lexical hypothesis: Identifying personality structure in natural language., Journal of Personality and Social Psychology (2022).
[16] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, S. Ravi, Goemotions: A dataset of fine-grained emotions, arXiv preprint arXiv:2005.00547 (2020).
[17] L. Feldman Barrett, J. A. Russell, Independence and bipolarity in the structure of current affect., Journal of personality and social psychology 74 (1998) 967.
[18] J. B. Nezlek, P. Kuppens, Regulating positive and negative emotions in daily life, Journal of personality 76 (2008) 561–580.
[19] M. B. Donnellan, F. L. Oswald, B. M. Baird, R. E. Lucas, The mini-ipip scales: tiny-yet-effective measures of the big five factors of personality., Psychological assessment 18 (2006) 192.
[20] L. A. Jensen-Campbell, W. G. Graziano, Agreeableness as a moderator of interpersonal conflict, Journal of personality 69 (2001) 323–362.
[21] B. W. Roberts, C. Lejuez, R. F. Krueger, J. M. Richards, P. L. Hill, What is conscientiousness and how can it be assessed?, Developmental psychology 50 (2014) 1315.
[22] L. D. Smillie, C. G. DeYoung, P. J. Hall, Clarifying the relation between extraversion and positive affect, Journal of personality 83 (2015) 564–574.
[23] S. Balta, E. Emirtekin, K. Kircaburun, M. D. Griffiths, Neuroticism, trait fear of missing out, and phubbing: The mediating role of state fear of missing out and problematic instagram use, International Journal of Mental Health and Addiction 18 (2020) 628–639.
[24] R. R. McCrae, D. M. Greenberg, Openness to experience, The Wiley handbook of genius (2014) 222–243.
[25] R. A. Martin, P. Puhlik-Doris, G. Larsen, J. Gray, K. Weir, Individual differences in uses of humor and their relation to psychological well-being: Development of the humor styles questionnaire, Journal of research in personality 37 (2003) 48–75.
[26] H. Medland, K. De France, T. Hollenstein, D. Mussoff, P. Koval, Regulating emotion systems in everyday life, European Journal of Psychological Assessment (2020).
[27] A. C. Schat, E. K. Kelloway, S. Desmarais, The physical health questionnaire (phq): construct validation of a self-report scale of somatic symptoms., Journal of occupational health psychology 10 (2005) 363.
[28] A. Kokoszka, A. Jastrzębski, M. Obrębski, Ocena psychometrycznych właściwości polskiej wersji kwestionariusza zdrowia pacjenta-9 dla osób dorosłych, Psychiatria 13 (2016) 187–193.
[29] D. Preece, R. Becerra, K. Robinson, J. Dandy, A. Allan, The psychometric assessment of alexithymia: Development and validation of the perth alexithymia questionnaire, Personality and Individual Differences 132 (2018) 32–44.
[30] R. Likert, A technique for the measurement of attitudes., Archives of psychology (1932).
[31] S. Cohen, R. C. Kessler, L. U. Gordon, Measuring stress: A guide for health and social scientists, Oxford University Press on Demand, 1997.
[32] E. Diener, D. Wirtz, W. Tov, C. Kim-Prieto, D.-w. Choi, S. Oishi, R. Biswas-Diener, New well-being measures: Short scales to assess flourishing and positive and negative feelings, Social indicators research 97 (2010) 143–156.
[33] E. Diener, R. A. Emmons, R. J. Larsen, S. Griffin, The satisfaction with life scale, Journal of personality assessment 49 (1985) 71–75.
[34] K. Förster, P. Kanske, Upregulating positive affect through compassion: Psychological and physiological evidence, International Journal of Psychophysiology 176 (2022) 100–107.
[35] G. Haydon, J. Reis, L. Bowen, The use of humour in nursing education: An integrative review of research literature, Nurse Education Today (2023) 105827.
[36] K. Pearson, Vii. note on regression and inheritance in the case of two parents, proceedings of the royal society of London 58 (1895) 240–242.

A. Annotator Profiles

Annotator profiles comprised with the results of the questionnaires mentioned in 3.3 are presented in Fig. 8.

Personality: Agreeableness: The average score of 15.6 suggests that people tend to be moderately cooperative and compassionate towards others (with a standard deviation of 2.2). Extraversion: The average score of 12.1 indicates that, on average, individuals tend to have a moderate level of sociability and assertiveness (with a standard deviation of 4.3). Conscientiousness: With an average score of 15.1, individuals, on average, exhibit a moderate level of organization and responsibility (with a standard deviation of 2.7). Neuroticism: The average score of 12.8 implies that, on average, individuals tend to have a moderate level of emotional stability and experience negative emotions (with a standard deviation of 3.6). Intellect: The average score of 15.1 suggests that, on average, individuals tend to exhibit a moderate level of intellectual curiosity and openness to new ideas (with a standard deviation of 2.5). **Humor Style:** Affiliative humor: The average score of 29.4 indicates that on average people tend to use humor extensively to strengthen social bonds and improve relationships (with a standard deviation of 5.5). Self-enhancing humor: The average

score of 25.2 suggests that, on average, individuals tend to use humor extensively as a coping mechanism to maintain a positive outlook during stressful situations (with a standard deviation of 6.7). Aggressive humor: With an average score of 19.5, individuals, on average, exhibit a moderate tendency to use humor as a means of teasing or mocking others (with a standard deviation of 4.8). Self-defeating humor: The average score of 19.1 implies that, on average, individuals tend to moderately engage in self-disparaging humor and put themselves down (with a standard deviation of 5.4). **Stress and Emotions:** Alexithymia: The average score of 14.8 indicates that, on average, individuals tend to have a low level of difficulty in identifying and expressing emotions (with a standard deviation of 6.4). Stress: With an average score of 14.5, individuals, on average, perceive a low level of stress in their lives (with a standard deviation of 6.8). Positive affect: The average score of 22.4 suggests that, on average, individuals experience a moderate level of positive emotions (with a standard deviation of 4.7). Negative affect: The average score of 17.5 implies that, on average, individuals experience a moderate level of negative emotions (with a standard deviation of 5.6). Satisfaction with life: The average score of 21.6 indicates that, on average, individuals have a low level of satisfaction and happiness with their lives (with a standard deviation of 5.7). **Emotion's Regulation:** Relaxation: The average score of 94.4 suggests that, on average, individuals engage in relaxation techniques to manage their emotions to a moderate extent (with a standard deviation of 61.7). Engagement: With an average score of 125.5, individuals, on average, exhibit a moderate level of involvement and immersion in activities as a means of emotion regulation (with a standard deviation of 62). Rumination: The mean score is 93.7, reflecting a moderate tendency to ruminate or dwell on negative thoughts or emotions (with a standard deviation of 65.3). Reappraisal: The mean score is 115.5, indicating a moderate tendency to reinterpret situations to regulate emotions (with a standard deviation of 58.1). Distraction: The mean score is 89.6, reflecting a moderate preference for using distractions as an emotion regulation strategy (with a standard deviation of 63.4). Suppression: The mean score is 46.5, indicating a relatively lower tendency to suppress or hide emotions (with a standard deviation of 50.4). **Health** Depression: The average score of 6.1 indicates that, on average, individuals report a relatively low level of depression (with a standard deviation of 5.6). Sleep disturbance: The average score of 10.7 suggests that, on average, individuals experience a low level of sleep disturbance (with a standard deviation of 6.0). Headaches: The average score of 6.7 indicates that, on average, individuals report a low level of headaches (with a standard deviation of 4.3). Gastrointestinal Problems: The average score of 7.4 suggests that, on average, individuals report a low level of gastrointestinal problems (with a standard deviation of 4.3). Respiratory Infections: The average score of 3.5 indicates that, on average, individuals report a relatively low level of respiratory infections (with a standard deviation of 2.7).

B. Heatmaps

Heatmaps may vary in the number of dimensions displayed in the scope of biases and the results of the questionnaire. Only dimensions that exhibit a correlation value of 0.1 or higher are displayed.

B.1. Personality Traits

In Fig. 9, **agreeableness** is moderately correlated with the dimension connected with positive affect dimensions (*positive*, *delight*, *inspiration*, *surprise* and *compassion*) whereas weakly with *joy*. The relationship with negative affect dimensions is slightly weaker. Data analysis revealed a weak positive relationship between **extraversion** and selected positive emotion bias. There is a weak (positive, surprise and compassion) and moderate (delight, inspiration, joy) relationship between **conscientiousness** trait and a few positive affect biases, and a weak negative relationship between negative emotion bias (negative, sadness, and anger). Two rational bias (*offensive to me* and *funny to me*) are related to conscientiousness. There is a weak negative correlation between **neuroticism** and positive affect biases (*joy*, *delight*, *inspiration* and *compassion*). A similar relationship is observed for some negative affect biases (*negative*, *fear*) and the opposite for *anger*. This trait is positively weekly correlated with rational biases (*embarrassing*, *vulgar*, *political*, *understandable*, *offensive to someone*). The inverse relationship is observed for *anger* and *offensive* to me biases. **Intellect** is negatively related to three rational biases (*offensive to me*, *vulgar* and *embarassing*) and positively with two (*political* and *understandable*). There is a weak negative association between intellect and positive affect biases (*compassion*, *surprise*, *calm*, *inspiration* and *delight*).

B.2. Humor

In Fig. 10, **affiliative humor** has a weak positive correlation with rational dimension biases (*understandable*) and a negative correlation with *embarrassing* and *interesting*. For the negative affect dimension biases, a weak positive correlation can be seen for *negative*, *fear*, *sadness*, *disgust* and *anger* bias. **Self-enhancing humor** is negatively correlated with rational dimension biases (funny to me, offensive to someone, understandable, interesting, political, embarrassing). **Aggressive humor** is positively cor-

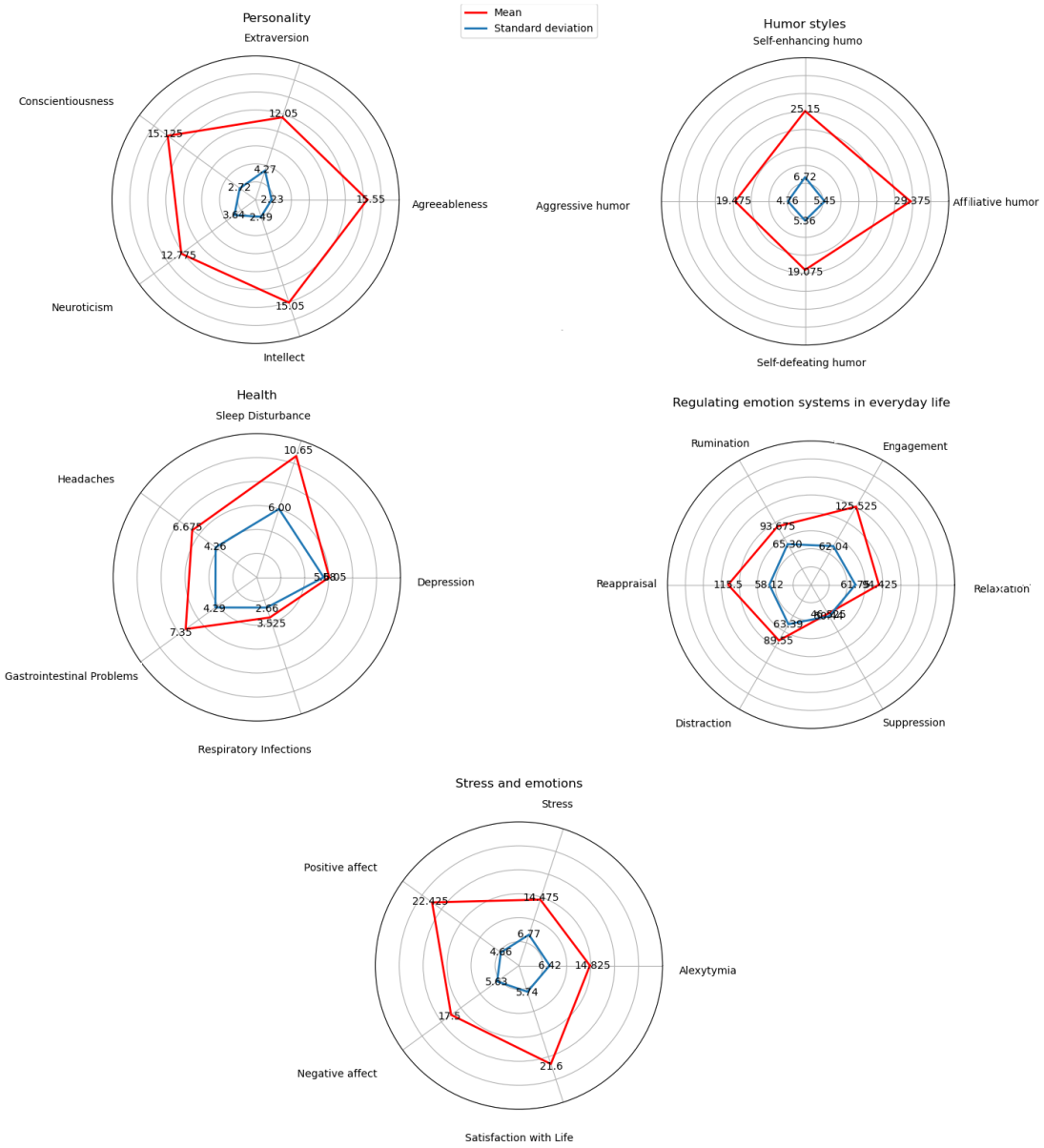

Figure 8: Detailed characteristics of the annotators

related with negative affect dimensions (*funny to someone*, *offensive to someone*, *understandable*,*interesting* and *political*) and positive dimensions biases (*positive*, *joy*, *delight*, *surprise*) and negative dimension bias (disgust). **Self-defeating humor** is correlated with positive affect dimensions (*positive*, *joy*, *delight*, *inspiration*, *surprise*, *compassion*), negative affect dimensions (*negative*, *fear*, *sadness*, *disgust* and *anger*) and rational dimensions (*embarrassing*, *interesting*) and negatively with (*political*, *understandable*).

B.3. Emotion Regulation

In Fig. 11, **engagement** has only weak negative correlations with the rational dimension (*ironic, embarrassing, vulgar, understandable, offensive to someone, funny*

Figure 9: Relationship between personality traits and annotation dimensions

to someone). For the positive affect dimension, a weak positive correlation can be seen for *positive* and *calm* bias. There is also a weak positive correlation for *negative* bias, and a weak negative correlation is for *disgust* biases.

There is a moderate positive relationship between *offensive to someone* bias and **rumination**. For the other rational dimension (*funny to someone, funny to me*) a weak positive correlation occurs. The correlations for positive affects dimension are similarly distributed: moderate for surprise bias and weak positive for *compassion, positive, calm and inspiration* bias.

Reappraisal is weakly correlated with the dimension associated with positive dimensions (*surprise, positive, compassion*) and negative dimensions (*negative, sadness*). The same absolute value occurs for *ironic* and *funny to me bias*, except that the former shows a weak positive correlation and the latter a weak negative correlation.

There is a moderate relationship between **distraction**

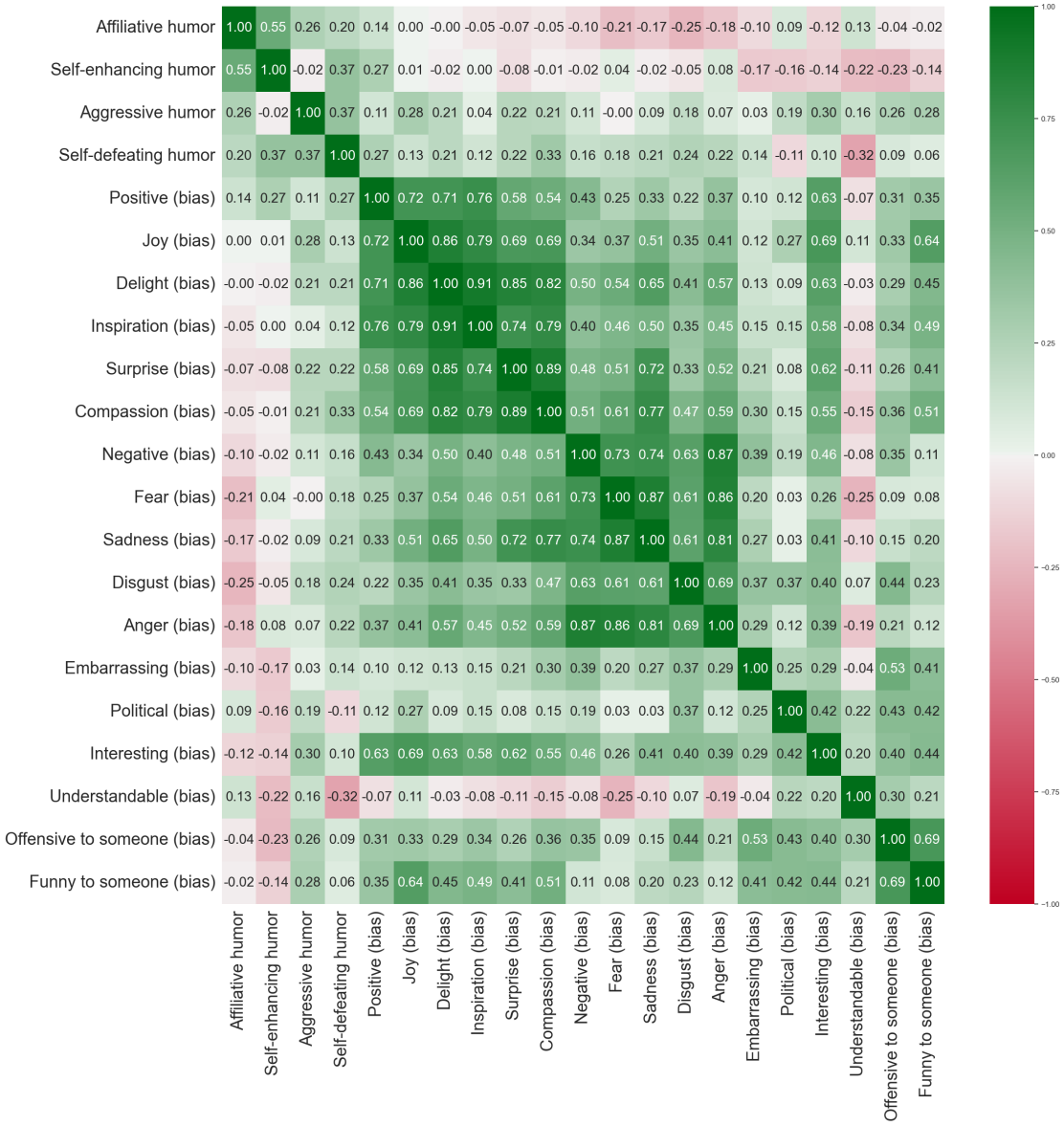

Figure 10: Relationship between humor and annotation dimensions

and four positive dimensions (*positive, inspiration, joy, compassion*). A similar correlation occurs for the two rational dimensions (*offensive to someone, funny to someone*). Other rational (*offensive to me, funny to me, ironic, vulgar*) and positive (*delight, surprise*) dimensions show a weak positive correlation.

Suppression is moderately correlated with two rational dimension (*funny to someone, funny to me*). Other rational dimension (*vulgar, embarrassing, offensive to someone, ironic*) have a weak positive relationship. A similar correlation is found with *compassion* bias.

B.4. Health

In Fig. 12, there is a weak relationship between **depression** and disgust bias. At the same time, there is a negative correlation with positive affects (joy, positive). From positive affects only compassion is related to depression in a weak positive correlation. There are also relationships with rational differentiation: weak positive (ironic,

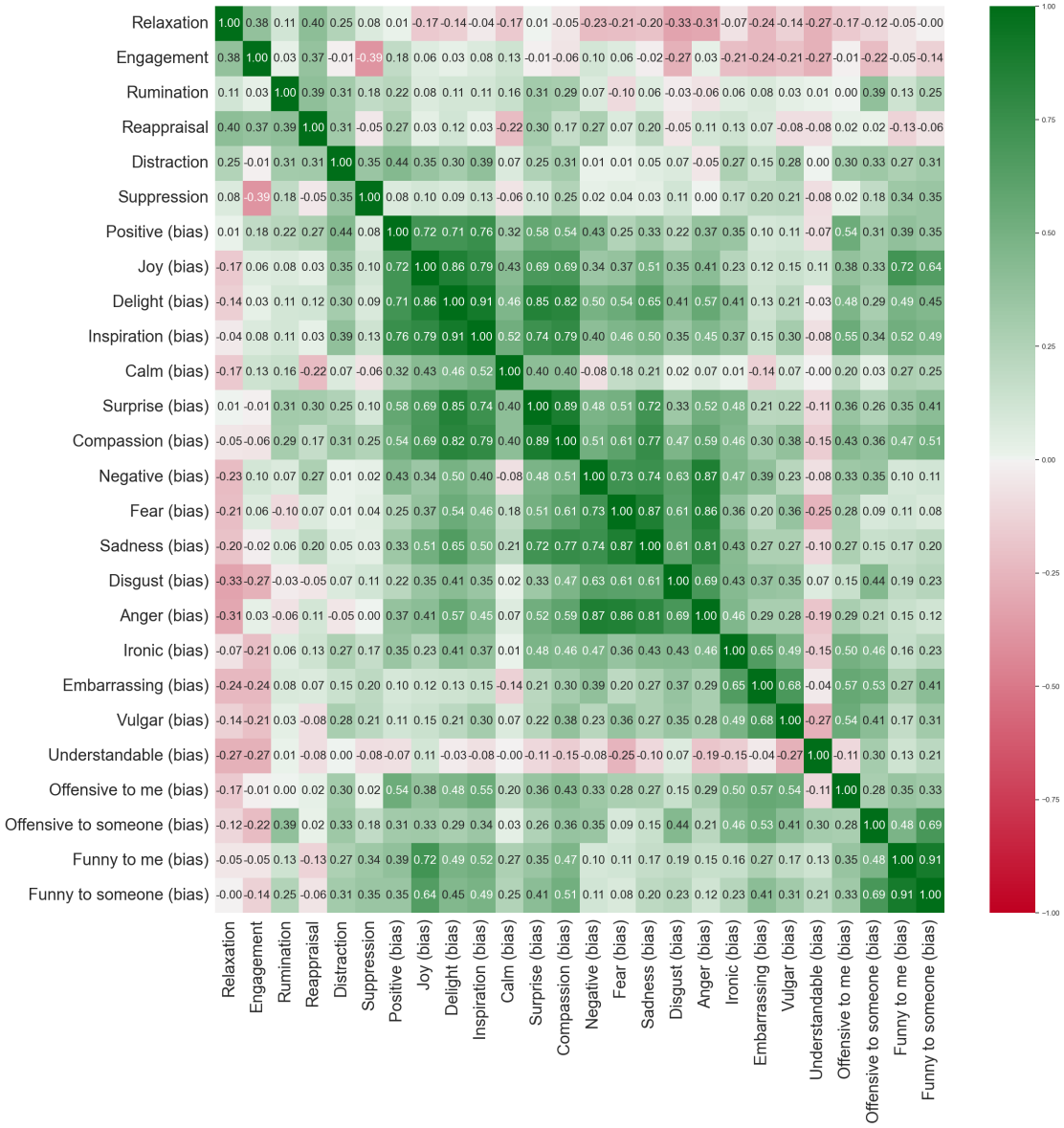

Figure 11: Relationship between emotion's regulation strategy and annotation dimensions

offensive to someone), moderate positive (vulgar, embarrassing), and moderate negative correlation (understandable).

A positive affect (*compassion*) is weakly positively related to **sleep disturbance**. There is a similar but negative correlation with *joy* bias and rational effects (*understandable, interesting, incomprehensible*). There are also weak positive correlations with rational affects (*embarrassing,vulgar*). The *ironic* bias has a weak positive correlation. There is no relationship between sleep disturbance and positive bias.

The item most highly correlated is understandable bias, with a moderate positive correlation with **headaches**. Other rational dimensions show weak positive correlations (*ironic, interesting, embarrassing, vulgar*). There is a weak positive association between headache and negative affect (*anger*). There is a weak negative correlation with incomprehensible bias.

Gastrointestinal problems are mostly correlated with rational affect: with moderate positive correlation

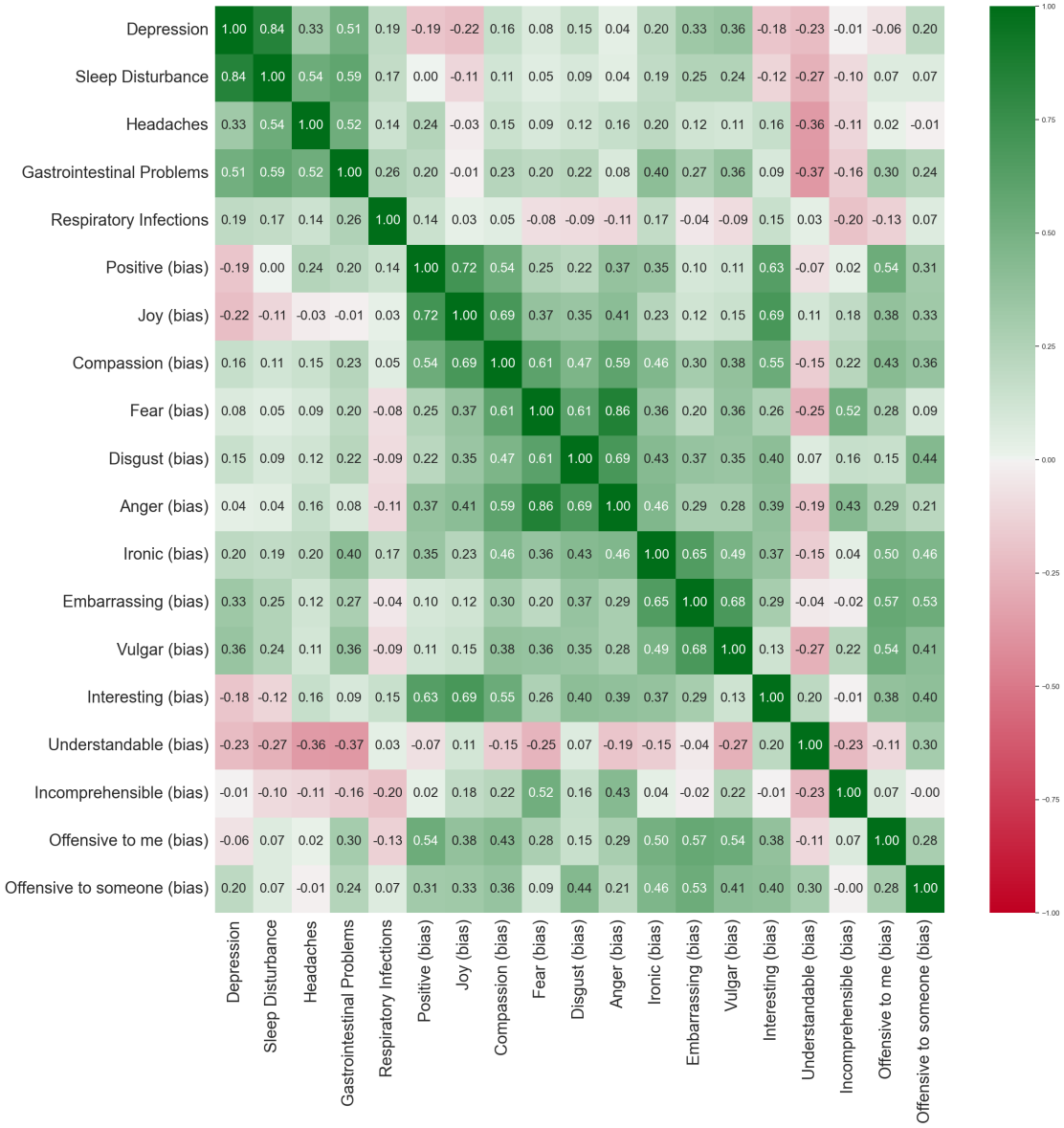

Figure 12: Relationship between depression and health with annotation dimensions

(*ironic, vulgar, offensive to me*), weak positive correlation (*embarrassing, offensive to someone*) and negative correlations: moderate (*understandable*) and weak (*incomprehensible*). There are as many weak positive relationships with positive affects (*compassion, positive*) as with negative ones (*disgust, fear*).

For **respiratory infections**, the strongest correlations are with rational affect, both weakly positive (*ironic, interesting*) and weakly negative (*incomprehensible, offensive to me*). Negative affect (*anger*) has a weak negative correlation. For positive bias, there is a weak positive correlation.

A moderate or weak negative correlation can be observed between perceiving a text as understandable and headaches, gastrointestinal problems, depression, and sleep disturbance. The same somatic health dimensions have a positive correlation with interpreting a text as vulgar or embarrassing. There is a positive correlation with the *ironic* bias in all physical health dimensions studied.

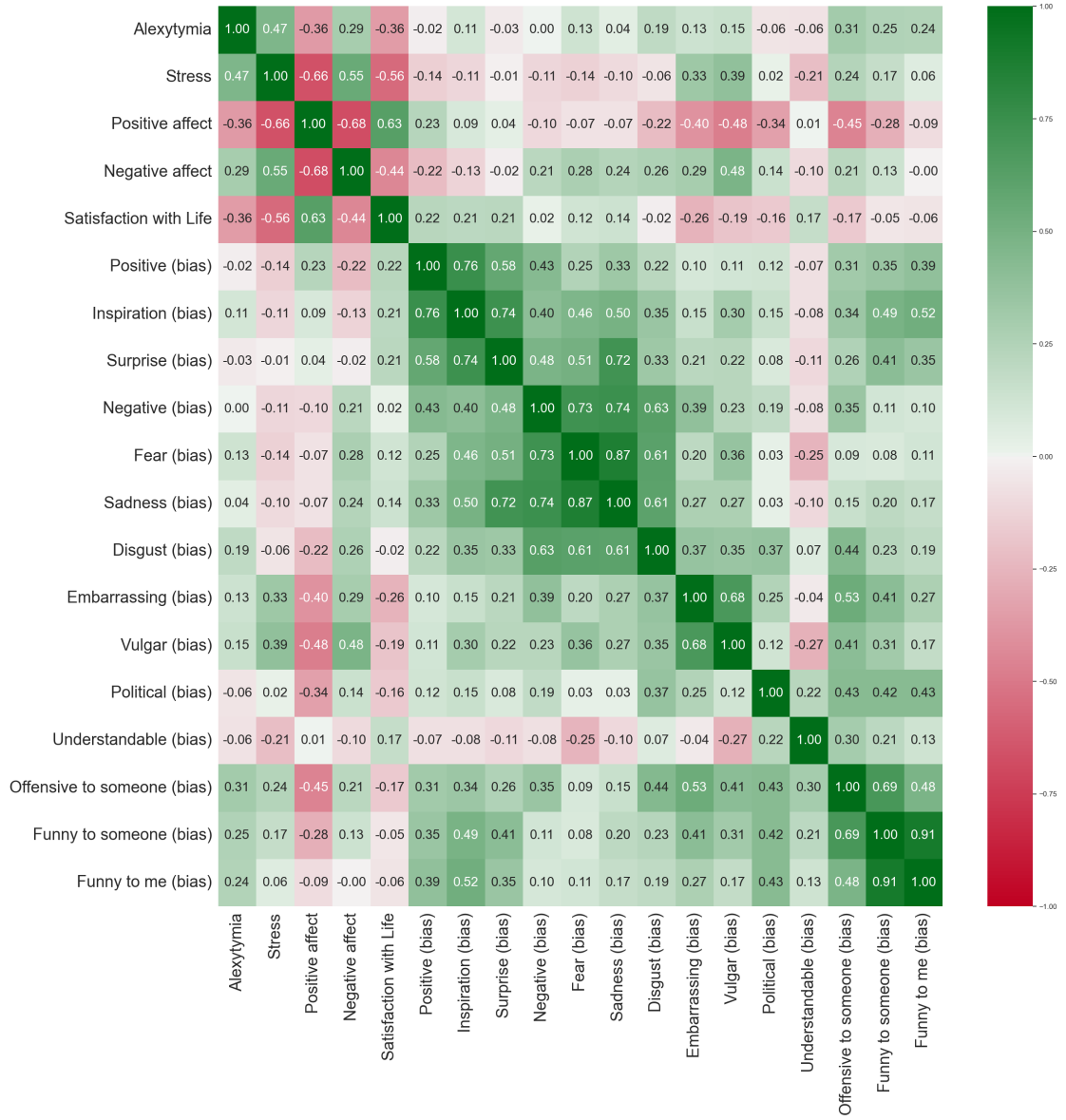

Figure 13: Relationship between stress, emotions, and annotation dimensions

B.5. Stress and Emotions

In Fig. 13, experiencing **stress** is moderate (*vulgar*, *embarrassing*) and weakly (*offensive to someone*, *funny to someone*) positively related to rational biases. The negative relationship is noticed only with *understandable* rational bias. **Positive affect** (*positive*, *inspiration*) and **negative affect** (*negative*, *fear*, sadness) biases are weakly negative related to stress. Experienced positive affect is negatively moderately related to rational biases (*funny to someone*, *offensive to someone*, *political*, *vulgar* and *embarassing*). Among positive affect dimensions only *positive* bias is positively correlated. From the negative affect dimensions, only disgust bias is weakly negatively correlated. Both the negative affect dimensions (*negative*, *fear*, *sadness*, *disgust*) and rational dimensions (*embarrassing*, political, *offensive to someone* and *funny to someone*) are weakly related to negative affect. Only the *vulgar* bias is moderately related to the negative affect. There is a weak positive relationship between positive dimensions

biases (*positive*, *inspiration* and *surprise*) and **satisfaction with life**. There is a weak inverse relationship between negative dimension biases (*fear*, *sadness*). Rational dimension biases (*embarrassing*, *vulgar*, *political*, *offensive to someone*) are weakly negatively related to satisfaction with life. A positive correlation is only observed for the *understandable* bias.

C. Intra-Annotator Agreement

C.1. Intra-Annotator Agreement for Affective Dimensions

Tab. 4 provides information about the Intra-Annotator Agreement for 3 selected users: strict IntraAA, strict IntraAA measured without consistent zero labels, and soft IntraAA measures for the affective dimensions. There is also information on the proportion of zero-markings, which can provide a reference point for interpreting particular results.

The increase/decrease percentages highlight how the scores change when moving from strict intraAA to soft intraAA. It may be assumed that the larger the increase, the less calibrated the judgment within a given dimension.

The Average results refer consecutively to the average: Strict IntraAA, Soft IntraAA, and Increase when switching from Strict to Soft.

C.2. Intra-Annotator Agreement for Rational Dimensions

Tab. 5 presents information similar to Tab. 4, but for rational dimensions.

Interestingly, we note here that there are significantly greater differences between the various dimensions in the level of increase when a soft IntraAA is used. The dimensions for which this increase was the smallest are shown here. They refer to offensiveness and funniness. This indicates that perception of these dimensions is the most formed and is least influenced by external circumstances.

Table 4
Intra-Annotator Agreement for affective dimensions (selected users)

	Positive	Negative	Joy	Delight	Inspiration	Calm	Surprise	Compassion	Fear	Sadness	Disgust	Anger
(A)												
Strict IntraAA (user 0)	0,41	0,47	0,71	0,65	0,55	0,39	0,36	0,59	0,53	0,69	0,67	0,65
Strict IntraAA (user 0) [non-zero]	0,12	0,22	0,2	0,08	0,1	0,08	0,18	0,15	0,15	0,06	0,14	0,11
Proportion of null markings [user 0]	32,99%	32,99%	63,92%	61,86%	49,48%	34,02%	21,65%	51,55%	44,33%	67,01%	61,86%	60,82%
Strict IntraAA (user 10)	0,55	0,35	0,64	0,67	0,73	0,8	0,28	0,51	0,74	0,41	0,55	0,46
Strict IntraAA (user 10) [non-zero]	0,09	0,15	0,18	0,15	0,05	0,13	0,19	0,18	0,14	0,17	0,06	0,12
Proportion of null markings [user 10]	50,71%	23,57%	56,43%	61,43%	71,43%	77,14%	10,71%	40,00%	70,00%	28,57%	52,14%	38,57%
Strict IntraAA (user 15)	0,8	0,43	0,91	0,94	0,87	0,86	0,96	0,86	0,75	0,76	0,68	0,49
Strict IntraAA (user 15) [non-zero]	0,12	0,2	0,04	0,05	0,12	0,07	0	0,07	0,1	0,04	0,08	0,15
Proportion of null markings [user 15]	77,03%	28,62%	90,81%	93,29%	85,51%	84,81%	96,47%	84,81%	71,73%	75,27%	65,37%	39,93%
Average	0,59	0,42	0,76	0,75	0,72	0,68	0,53	0,65	0,67	0,62	0,63	0,53
(B)												
Soft IntraAA [user 0]	0,86	0,79	0,89	0,88	0,86	0,8	0,84	0,88	0,84	0,86	0,95	0,89
Soft IntraAA [user 10]	0,91	0,8	0,94	0,89	0,92	0,96	0,78	0,77	0,88	0,74	0,81	0,78
Soft IntraAA [user 15]	0,98	0,95	0,99	0,99	0,98	0,97	1	0,98	0,95	0,94	0,95	0,94
Average	0,91	0,85	0,94	0,92	0,92	0,91	0,87	0,87	0,89	0,84	0,9	0,87
Increase (A) ->(B)												
user 0	107,50%	67,39%	24,64%	34,92%	56,60%	105,26%	131,43%	49,12%	58,82%	23,88%	41,54%	36,51%
user 10	64,94%	128,57%	46,67%	32,98%	26,47%	19,64%	179,49%	52,11%	18,27%	80,70%	48,05%	70,31%
user 15	22,12%	121,49%	8,14%	5,28%	11,74%	12,76%	3,66%	13,58%	27,49%	22,69%	39,38%	92,09%
Average	64,85%	105,82%	26,48%	24,39%	31,61%	45,89%	104,86%	38,27%	34,86%	42,42%	42,99%	66,30%

Table 5
Intra-Annotator Agreement for rational dimensions (selected users)

	Ironic	Embarrassing	Vulgar	Political	Interesting	Understandable	Agreement	Trust	Incomprehensible	Sympathy	Offensive to me	Offensive to someone	Funny to me	Funny to Someone
(A)														
Strict IntraAA (user 0)	0,77	0,82	0,79	0,81	0,28	0,48	0,57	0,51	0,46	0,36	0,96	0,77	0,93	0,91
Strict IntraAA (user 0) [non-zero]	0,15	0,23	0,17	0,25	0,26	0,48	0,54	0,47	0,04	0,34	0,2	0,35	0,13	0,18
Proportion of null markings [user 0]	73,20%	77,32%	75,26%	75,26%	3,09%	0,00%	6,19%	6,19%	44,33%	3,09%	94,85%	64,95%	91,75%	88,66%
Strict IntraAA (user 10)	0,38	0,64	0,89	0,91	0,12	0,24	0,6	0,53	0,24	0,49	0,97	0,76	0,94	0,95
Strict IntraAA (user 10) [non-zero]	0,04	0,04	0,06	0	0,1	0,23	0,6	0,53	0,18	0,49	0,2	0,23	0,11	0,22
Proportion of null markings [user 10]	35,00%	62,86%	87,86%	91,43%	2,14%	1,43%	0,71%	0,00%	6,43%	1,43%	96,43%	68,57%	93,57%	93,57%
Strict IntraAA (user 15)	0,74	0,79	0,95	0,85	0,92	0,76	0,43	0,31	0,67	0,61	1	0,51	0,97	0,91
Strict IntraAA (user 15) [non-zero]	0,12	0,12	0,24	0,16	0,04	0,76	0,36	0,2	0,08	0,22	0	0,05	0	0
Proportion of null markings [user 15]	70,67%	76,33%	93,99%	82,33%	91,52%	0,00%	11,66%	13,43%	63,96%	50,18%	99,65%	48,41%	96,82%	90,81%
Average	0,63	0,75	0,88	0,86	0,44	0,5	0,53	0,45	0,46	0,49	0,98	0,68	0,95	0,92
(B)														
Soft IntraAA [user 0]	0,92	0,94	0,96	0,93	0,78	0,8	0,78	0,72	0,79	0,73	0,99	0,93	0,96	0,94
Soft IntraAA [user 10]	0,74	0,85	0,96	0,98	0,68	0,74	0,89	0,85	0,82	0,85	0,99	0,92	0,99	0,99
Soft IntraAA [user 15]	0,96	0,98	1	0,95	0,99	0,89	0,87	0,77	0,9	0,96	1	0,88	1	1
Average	0,87	0,92	0,97	0,95	0,82	0,81	0,84	0,78	0,84	0,85	0,99	0,91	0,98	0,98
Increase (A) ->(B)														
user 0	18,67%	13,75%	20,78%	13,92%	181,48%	65,96%	38,18%	42,86%	71,11%	102,86%	3,23%	20,00%	3,33%	3,41%
user 10	94,34%	32,22%	8,87%	7,03%	458,82%	205,88%	47,62%	60,81%	248,48%	72,46%	1,47%	21,70%	5,30%	4,51%
user 15	29,52%	23,21%	4,81%	12,03%	8,08%	16,67%	100,82%	150,57%	34,39%	57,23%	0,35%	72,22%	3,28%	10,12%
Average	47,51%	23,06%	11,49%	11,00%	216,13%	96,17%	62,21%	84,75%	118,00%	77,52%	1,68%	37,97%	3,97%	6,01%