# Emotion Ratings: How Intensity, Annotation Confidence and Agreements are Entangled

**Enrica Troiano, Sebastian Padó** and **Roman Klinger**
Institut für Maschinelle Sprachverarbeitung
University of Stuttgart, Germany
{firstname.lastname}@ims.uni-stuttgart.de

## Abstract

When humans judge the affective content of texts, they also implicitly assess the correctness of such judgment, that is, their confidence. We hypothesize that people's (in)confidence that they performed well in an annotation task leads to (dis)agreements among each other. If this is true, confidence may serve as a diagnostic tool for systematic differences in annotations. To probe our assumption, we conduct a study on a subset of the Corpus of Contemporary American English, in which we ask raters to distinguish *neutral* sentences from *emotion*-bearing ones, while scoring the confidence of their answers. Confidence turns out to approximate inter-annotator disagreements. Further, we find that confidence is correlated to emotion intensity: perceiving stronger affect in text prompts annotators to more certain classification performances. This insight is relevant for modelling studies of intensity, as it opens the question wether automatic regressors or classifiers actually predict intensity, or rather human's self-perceived confidence.

## 1 Introduction

A plethora of theories exist on the matter of emotions: the intensity of affective states, their link to cognition, and their arrangement into categories are just a few of the angles from which psychology has tackled this complex phenomenon (Gendron and Feldman Barrett, 2009). Correspondingly, in computational emotion analysis, texts have been associated to values of intensity (Strapparava and Mihalcea, 2007; Mohammad and Bravo-Marquez, 2017), to cognitive components (Hofmann et al., 2020), and discrete classes (Zhang et al., 2018; Zhong et al., 2019, i.a.). In support of these tasks, substantial research effort has been directed to resource construction, which typically relies on the participation of human judges. Yet, emotions are a subjective experience. Their interpretation in text varies across individuals, and this poses a major challenge in emotion analysis: it is hard to reach acceptable levels of inter-annotator agreement (IAA) (Bobicev and Sokolova, 2017; Schuff et al., 2017; Troiano et al., 2019).

On their side, humans (roughly) know how well they can "read" emotions (Realo et al., 2003). They judge the affective content of texts, and at the same time, the correctness of such judgment: in other words, annotators can evaluate their own confidence with respect to their labelling decisions. This hints a possible relation between confidence and IAA. One could expect that annotators are more likely to incur inconsistencies when they feel uncertain about their answers. Hence, it would be interesting to verify if self-assessed confidence approximates inter-annotator disagreements in affect-related annotation tasks. The collection of this type of judgments is not a common practice: past research has found that self-assigned scores of confidence are predictable based on some vocal attributes of emotion speech stimuli (Lausen and Hammerschmidt, 2020), but this has not been done on text, to the best of our knowledge.

Yet another aspect involved in emotion recognition and which, at first sight, relates to confidence, is emotion intensity. It would be intuitive to assume that emotions are recognized with higher confidence if they are expressed with stronger magnitude (e.g., "*The teacher exploded*" > "*He snapped his annoyed temper*", "*sadder*" > "*a bit sad*", "*ecstasy*" > "*joy*"). Still, in the sentence "*We had to cheer him up; later, he was off the ground*", readers have to choose what part of the text to attend to (the challenge that the speakers undertook – not intense, or the effect they had – intense) and be very confident about either choice. Similar counter examples reveal that the link between the perception of emotion intensity and the self-perceived confidence is opaque, and leaves room for exploration.



*Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 40–49

In this paper, we experimentally investigate the relationship between three human judgments: about the presence of emotions, about their intensity, and about the confidence of the annotation decision. Leveraging such information, we aim at understanding in what cases annotators differ regarding the judgement that an emotion is expressed. Our first research question is: (RQ1) Are (dis)agreements with respect to the presence of emotions related to confidence? Second, (RQ2) Are judgments of intensity and confidence entangled? We bring these issues together in an annotation study based on emotion recognition, in which self-perceived confidence is a dimension to be rated. Given a subset of the Corpus of Contemporary American English (COCA) (Davies, 2015), raters distinguish emotion-bearing sentences from neutral ones, while quantifying both the intensity of the emotion and their confidence on a Likert scale (Bègue et al., 2018).

We find that confidence can explain systematic differences in decisions of annotators; so does intensity; and impressions on intensity and confidence are correlated. Based on these results, we devise a strategy to leverage the two factors and smoothen out inter-annotator inconsistencies.

## 2 Annotation Setup

**Tasks.** The first step in this study is to collect emotion assessments.[1] We are not interested in which emotion people interpret from text, but rather if they recognize any. Judges read sentences and answer the question: *(*EMO*) Is it Emotional or Neutral?* For the items deemed to express an emotion, we also ask *(*INT*) How strong is it?*, which enables us to obtain ratings about affective strengths on a Likert scale from 1 (not intense) to 3 (very). Lastly, since raters interpret emotions without an immediate first-hand experience, we have them self-evaluate their own judgments on a scale from 1 (unsure) to 3 (certain), in response to the question *(*CONF*) How confident are you about your answer to* EMO*?*

As for EMO, we acknowledge that the emotional content of an utterance can be inferred from many perspectives. It is possible to assess one's own emotion after reading the text, to reconstruct the affective state of the writers who produced it, to guess the reaction that they intended to elicit in the readers, and so on. To avoid confusion, we instruct annotators to consider the presence of an emotion only with respect to their personal viewpoint.

We opt for an in-lab setting. Raters are three female master students aged between 24 and 27, who are proficient in English, and have some annotation experience and background in computational emotion analysis.

**Data.** Corpora that include emotion classes or gradations are tailored on specific domains, like self-reports (Scherer and Wallbott, 1997), tweets (Mohammad, 2018) and newspapers (Strapparava and Mihalcea, 2008). We broaden our focus to multiple genres, and annotate sentences from the 2020 version of COCA[2], which includes unlabelled texts that occurred from 1990 to present in different domains, like blogs, magazines, newspapers, academic texts, spoken interactions, fictions, TV, and movie subtitles.

With a corpus of this size (>1B words), considering all data points would be costly, and randomly selecting them may cause imbalance in the final annotation – i.e., a majority of *neutral* instances. Therefore, we draw a sample biased towards emotional sentences with a combination of rules and classifier-based information. To obtain such a classifier, we fine-tune the pre-trained BERT (Devlin et al., 2019) base-case model on a number of emotion analysis resources[3], adding a classification layer that outputs the labels *emotion* or *neutral*. Having that, we filter academic texts out of COCA for their arguably impartial language, and from each of the other genres, we randomly pick 500 sentences; out of these, we sample 100 sentences balanced by class, i.e., 50 labelled as *neutral* by our classifier, 50 as bearing an *emotion*. Thus, the annotators are shown 700 items, 100 per domain, with a balanced class distribution, according to the classifier.

## 3 Results

To answer our research questions, we first observe annotators' agreements on EMO. The highest Cohen's $\kappa$ (1960) between pairs of human judges was .43 (Table 1a); Fleiss' $\kappa$ (1971) for the three annotators was .34.

At first glance, these numbers appear unsatisfactory. On the one hand, they are due to the skewed class distribution in the annotators' choices.[4] On

[1] The guidelines are in the Appendix. Our data is accessible at https://www.ims.uni-stuttgart.de/data/emotion-confidence.

[2] https://www.english-corpora.org/coca/

[3] Details and classifier's performance in Appendix.

[4] With skewed class distribution, chance agreement increases, penalizing the resulting $\kappa$ (Cicchetti and Feinstein, 1990).

| | IAA |
|---|---|
| A1–A2 | .38 |
| A2–A3 | .43 |
| A3–A1 | .30 |

(a) Cohen's $\kappa$ for annotator pairs on EMO.

| | Counts | |
|---|---|---|
| | 1 vs. 2 | 3 vs. 0 |
| E | 138 | 304 |
| N | 170 | 88 |

(b) Counts of *emotion* (E) and *neutral* (N) items for EMO answers, aggregated by agreement (1 vs. 2, 3 vs. 0).

| | CONF | INT |
|---|---|---|
| 1 | −.001 | .04 |
| 2 | .03 | .20 |
| 3 | .39 | .30 |

(c) Fleiss' $\kappa$ on EMO for each value of CONF and INT.

Table 1: Inter-Annotator Agreements.

the other, they can be traced back to the way in which the EMO task was formulated: asking if a text is emotional from the readers' own point of view (e.g., it "describes an event [...] to which *you* would associate an emotion", see Appendix) as opposed, for instance, to the writers', paves the way for more heterogeneous responses.

However, a look at other IAA measures, like the absolute counts of items that were assigned to each label, leads to a more detailed picture. Table 1b breaks down the annotated categories by agreement: column 1 vs. 2 corresponds to the groups of items on which 1 annotator chose a label, while the majority opted for the other; column 3 vs. 0 shows how many times all three annotators agreed. We see that 138 sentences out of 700 were deemed emotionally charged by only 1 person (and hence, were associated to *neutral* by the two others). 2 annotators picked the *emotion* class for 170 sentences, i.e., those which were *neutral* according to just 1 rater. Overall, as the amount of considered judgments increases, so does their intersubjective validity about emotions. This tendency is clear in column 3 vs. 0, which shows that there were more emotional instances with 3 identical labels than those with conflicting ratings. Indeed, perfect agreement was reached for 392 items (304 *emotion* and 88 *neutral*), more than half the data, suggesting that people had a shared understanding.

### 3.1 Confidence Approximates Disagreements (and so does Intensity)

We next focus on the items that received incoherent judgments. Annotators seem to diverge on the presence/absence of emotions in a systematic way. Specifically, their inconsistencies correspond to certain patterns in the ratings of confidence, as well as intensity. Taking pairs of annotators, we see that the one who picks the *emotion* class tends to do that with low confidence.[5] As an example representative of the general trend, on 11 sentences annotator A3 makes the *neutral* choice, while annotator A1 picks *emotion*, but rates such items with confidence 1 (5 sentences), or 2 (6 sentences) – never using the highest degree of confidence. The same holds for intensity: A1 rates 10 of the 11 sentences as having the lowest intensity, and 1 sentence as having intensity 2 – none with the highest intensity value.

Hence, to answer RQ1, the evaluation of intensity and the self-evaluation of confidence underlie disagreements in discrete emotion annotations. Further, they show that different intuitions are not totally incompatible, since the annotator who takes the *emotion* decision does so without being extremely confident, and gauging intensity as rather weak.

We corroborate this finding by looking at a more standard measure of IAA, namely Fleiss' $\kappa$, which turns out to be affected by both CONF and INT. We compute it once more for the answers of EMO, but here we consider to be emotional only the items on which all annotators choose a certain level of confidence or intensity. Table 1c displays $\kappa$ separately for different levels (rows) of CONF and INT. Low values aside, these results inform us that the lower CONF/INT, the more prominent are disagreements. The highest IAA (.39), instead, is achieved for the most confident answers.

**Post-Processing Disagreements.** If systematic differences among annotators can be diagnosed with the help of confidence and intensity, can they also be resolved to some extent? We use the CONF and INT scores as acceptance thresholds for the label *emotion*, so to post-process the EMO decisions of each judge: they turn into *neutral* if the corresponding CONF or INT answer does not reach a certain threshold t. For instance, using INT as thres-

[5] Ratings on disagreements are in Appendix, Table 4.

| | EMO CONF<2 | | EMO CONF< | |
|---|---|---|---|---|
| | 1 vs. 2 | 3 vs. 0 | 1 vs. 2 | 3 vs. |
| E | 172 | 187 | 141 | |
| N | 169 | 172 | 73 | 4 |

| | EMO INT<2 | | EMO INT< | |
|---|---|---|---|---|
| | 1 vs. 2 | 3 vs. 0 | 1 vs. 2 | 3 vs. |
| E | 165 | 82 | 57 | |
| N | 118 | 335 | 17 | 6 |

Table 2: Counts of labels for subsets of ratings on EMO, post-processed with acceptance thresholds <2 and <3, for both CONF (top) and INT (bottom).

hold, with t<2 all items labeled *emotion* in EMO are kept as such only in case the INT is 2 or more, all the others are mapped to *neutral*.

Agreement counts on the post-processed annotation of EMO are in Table 2. We see, again, that the number of agreed upon items increases by increasing the sets of equal ratings. For instance, 283 sentences received 2 unanimous judgments (column 1 vs. 2, under EMO INT<2), and 417 received 3. In comparison to the original annotation in Table 1b, we can observe a considerable change in the number of items with perfect agreement. While in the raw judgments they were 392, with t<3 they increase to 626. We find a similar pattern when leveraging confidence: with t<2 (low confidence), it is obtained for 359 items, and with t<3 (moderate), perfect agreement increases to 486 items.

This comes at the cost of agreeing on fewer *emotion* sentences (304 before filtering, 7 and 56 after applying the highest threshold to INT and CONF), but it indicates that the better raters agree on intensity or confidence, the more they agree regarding the presence or absence of emotions.

### 3.2 Stronger Intensity, Higher Confidence

Having found that confidence and intensity have a similar relationship to disagreements, we move to analyzing how they link to one another. To address RQ2, we focus on the ratings of the 304 sentences with the unanimous *emotion* judgment. For them, we compute the intra-annotator correlation between the answers to INT and the corresponding ratings of CONF. A Spearman's $\rho$ (Spearman, 1904) of .5 for annotator A1, .58 for A2 and .64 for A3 (p-value <.05 for all) reveals a significant positive correlation between intensity and confidence. This suggests that people believe they correctly classified a text if they also perceived high emotion intensity.

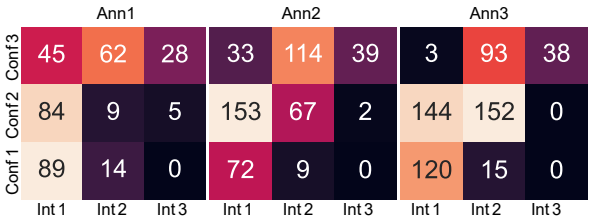


Figure 1: Cross-tabulation of INT and CONF by annotator, for the items that each of them deems emotional.

Figure 1 gives an in-depth account of the CONF-INT relation. It plots the counts of items that were labeled with a certain emotion intensity together with a certain confidence level, separately for each annotator. The columns INT3 tell us that rarely annotators perceive intensity as strong without being extremely confident that the text expresses an emotion. In fact, no instance was rated with the highest intensity and the lowest confidence (INT3-CONF1) at the same time. Conversely, for cases of low intensity, annotators tend to stay low also on the scale of confidence.

**On what do People Agree?** On the 304 sentences considered emotional by all, if the annotators gave the same score to intensity (i.e., perfect agreement is both on EMO and INT), they did not have a total disagreement on confidence (i.e., there are always at least 2 people with the same CONF) and vice versa. This may be a sign of the correlation between the two variables.

Moreover, some items were scored with perfect agreement on all questions. It is the case of sentences like "*I can't believe that you saved my life*", considered, with the highest level of confidence, to convey an emotion of intensity 3, "*Get off my back!*" deemed to have a mild intensity, and "*,,You have such an interesting life," she said, after a little small talk*" with intensity 1. While these examples are rated as highly certain (CONF3), there are also sentences on which people agree across all confidence-intensity combinations.[6] None of the instances has CONF1 and INT3, while a number of examples have CONF3 and INT1, indicating that it is harder to be uncertain of a strong emotion than to be sure of a weak one.

[6] A detailed analysis is in Table 5, Appendix.

## 4 Discussion and Conclusion

This annotation investigated if the perceived emotion (class), a perceived feature of emotion (intensity), and self-perception (confidence) are tied together – and can help understand inconsistent annotations. We found that (RQ1) both intensity and confidence account for inter-annotator inconsistencies relative to binary decisions. Adopting confidence as an acceptance threshold showed that higher scores lead to more uniform assessments of emotions; though not a surprising effect of confidence, this also applies to intensity. Moreover, (RQ2) the two variables are correlated, that is, people feel more certain about their emotion recognition performance on items with high intensity.

We acknowledge that our design of EMO naturally incurs the risk of inconsistent answers. However, precisely for the subjective nature of the task, the finding that disagreements decrease with high CONF/INT is interesting in itself: some judgments which are seemingly unsolvable can be explained by certain perceived properties of emotions (intensity) or self-perceived features (confidence).

From these results we can draw some lessons. First, the correlation between confidence and intensity brings relevant implications to all those studies that focus on emotional strength. When asked to evaluate intensity, do people confound that with confidence? Even more, is there a causal relation between the two? As a best practice to put safeguards in their guidelines, experimenters may ask people to tease the two variables apart. Potentially, this issue concerns modelling studies as well: do classifiers for emotion intensity predict such feature, or rather the confidence with which people judge emotions? We provided reasons to look into this further.

Second, confidence turned out to be an important dimension of rating, because it can inform us when the annotators expect to disagree. When judgments diverge, annotators do not deem their intuition credible. Hence, our finding that confidence approximates disagreements means precisely that people themselves predict their performance to differ from that of the others.

Concretely, all this knowledge can come in support of annotation studies. Including confidence as a rating dimension may give an additional source of information about annotators' reliability. This can help experimenters to refine the guidelines in a pre-testing phase: one might want to disentangle cases in which annotators' disagreements are random – signaling a lack of annotators' reliability, and when, instead, they are due to consistently different ways of perceiving and reporting on confidence. In this second case, disagreements may be normalized to an extent, by post-processing the annotation results. For instance, as people seem to agree on the class *emotion* only if they also agree on certain degrees of CONF, there might be some levels of confidence (or intensity) that one filters in/out of the final annotation labels.

A similar strategy may be somewhat restrictive, since it accepts as emotional only those items on which humans' intuition fall above a pre-defined threshold. While we observed agreement on such items, it is possible to adopt more nuanced evaluation approaches and integrate information about intensity or confidence into IAA measures. As an example, disagreements between two raters can be penalized more when the one choosing the *emotion* label does so by perceiving extreme confidence or intensity – even though we provided evidence that these cases are rare. Future work could explore this direction.

In summary, we uphold that disagreements are not necessarily symptomatic of unreliability. This claim has so far not found much attention in emotion annotations, but is in line with a more general body of research dedicated to the reasons and the patterns underlying annotators' disagreements and to the ways in which their intuitions should be aggregated and evaluated (Bayerl and Paul, 2011; Bhardwaj et al., 2010; Qing et al., 2014; Peldszus and Stede, 2013; Plank et al., 2014; Sommerauer et al., 2020, i.a.). The applicability of these ideas to emotions should not come as a surprise — their assessment can derive from perceptive and meta-perceptual phenomena (intensity and confidence, for instance). Therefore, if emotion judgments alone might not be sufficient to measure the quality of annotations, they can be enriched and, eventually, explained by the knowledge of such phenomena.

## Acknowledgements

This work was supported by Deutsche Forschungsgemeinschaft (project CEAT, KL 2869/1-2) and the Leibniz WissenschaftsCampus Tübingen "Cognitive Interfaces". We thank Laura Oberländer, Agnieszka Faleńska and the anonymous reviewers for their constructive feedback.

## References


Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 579–586, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Petra Saskia Bayerl and Karsten Ingmar Paul. 2011. What determines inter-coder agreement in manual annotations? a meta-analytic investigation. *Computational Linguistics*, 37(4):699–725.

Indrit Bègue, Maarten Vaessen, Jeremy Hofmeister, Marice Pereira, Sophie Schwartz, and Patrik Vuilleumier. 2018. Confidence of emotion expression recognition recruits brain regions outside the face perception network. *Social Cognitive and Affective Neuroscience*, 14(1):81–95.

Vikas Bhardwaj, Rebecca Passonneau, Ansaf Salleb-Aouissi, and Nancy Ide. 2010. Anveshan: A framework for analysis of multiple annotators' labeling behavior. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 47–55, Uppsala, Sweden. Association for Computational Linguistics.

Victoria Bobicev and Marina Sokolova. 2017. Inter-annotator agreement in sentiment analysis: Machine learning perspective. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 97–102, Varna, Bulgaria. INCOMA Ltd.

Laura-Ana-Maria Bostan and Roman Klinger. 2018. An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Domenic V Cicchetti and Alvan R Feinstein. 1990. High agreement but low kappa: Ii. resolving the paradoxes. *Journal of clinical epidemiology*, 43(6):551–558.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Mark Davies. 2015. Corpus of Contemporary American English (COCA).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378–382.

Maria Gendron and Lisa Feldman Barrett. 2009. Reconstructing the past: A century of ideas about emotion in psychology. *Emotion Review*, 1(4):316–339.

Diman Ghazi, Diana Inkpen, and Stan Szpakowicz. 2015. Detecting emotion stimuli in emotion-bearing sentences. In *Computational Linguistics and Intelligent Text Processing*, pages 152–165. Springer International Publishing.

Jan Hofmann, Enrica Troiano, Kai Sassenberg, and Roman Klinger. 2020. Appraisal theories for emotion classification in text. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 125–138, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Adi Lausen and Kurt Hammerschmidt. 2020. Emotion recognition and confidence ratings predicted by vocal stimulus type and prosodic parameters. *Humanities and Social Sciences Communications*, 7(1):1–17.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Vicki Liu, Carmen Banea, and Rada Mihalcea. 2007. Grounded emotions. In *Seventh International Conference on Affective Computing and Intelligent Interaction (ACII 2017)*, pages 477–483. IEEE Computer Society.

Saif Mohammad. 2012. #emotional tweets. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 246–255, Montréal, Canada. Association for Computational Linguistics.

Saif Mohammad. 2018. Word affect intensities. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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

Andreas Peldszus and Manfred Stede. 2013. Ranking the annotators: An agreement study on argumentation structure. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 196–204, Sofia, Bulgaria. Association for Computational Linguistics.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511, Baltimore, Maryland. Association for Computational Linguistics.

Ciyang Qing, Ulle Endriss, Raquel Fernández, and Justin Kruger. 2014. Empirical analysis of aggregation methods for collective annotation. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1533–1542, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Anu Realo, Jüri Allik, Aire Nõlvak, Raivo Valk, Tuuli Ruus, Monika Schmidt, and Tiina Eilola. 2003. Mind-reading ability: Beliefs and performance. *Journal of Research in Personality*, 37(5):420–445.

Klaus R. Scherer and Harald G. Wallbott. 1997. The ISEAR questionnaire and codebook. Geneva Emotion Research Group. https://www.unige.ch/cisa/research/materials-and-online-research/research-material/.

Hendrik Schuff, Jeremy Barnes, Julian Mohme, Sebastian Padó, and Roman Klinger. 2017. Annotation, modelling and analysis of fine-grained emotions on a stance and sentiment detection corpus. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 13–23, Copenhagen, Denmark. Association for Computational Linguistics.

Pia Sommerauer, Antske Fokkens, and Piek Vossen. 2020. Would you describe a leopard as yellow? evaluating crowd-annotations with justified and informative disagreement. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4798–4809, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Charles Spearman. 1904. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):71–101.

Carlo Strapparava and Rada Mihalcea. 2007. SemEval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74, Prague, Czech Republic. Association for Computational Linguistics.

Carlo Strapparava and Rada Mihalcea. 2008. Learning to identify emotions in text. In *Proceedings of the 2008 ACM symposium on Applied computing*, pages 1556–1560.

Enrica Troiano, Sebastian Padó, and Roman Klinger. 2019. Crowdsourcing and validating event-focused emotion corpora for German and English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4005–4011, Florence, Italy. Association for Computational Linguistics.

Yuxiang Zhang, Jiamei Fu, Dongyu She, Ying Zhang, Senzhang Wang, and Jufeng Yang. 2018. Text emotion distribution learning via multi-task convolutional neural network. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4595–4601. International Joint Conferences on Artificial Intelligence Organization.

Peixiang Zhong, Di Wang, and Chunyan Miao. 2019. Knowledge-enriched transformer for emotion detection in textual conversations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 165–176, Hong Kong, China. Association for Computational Linguistics.

## A Annotation Guidelines

In what follows are the detailed guidelines of the annotation. Q1 corresponds to EMO in the paper, Q2 corresponds to CONF, and Q3 to INT.

In Q1, annotators were required to disregard the emotion that the author of the utterances intended to express or to elicit in others. Their task was to give their immediate, personal impression with respect to the presence of an emotion.

Before annotating the 700 sentences selected from COCA, we observe inter-annotator agreement on a pre-annotation trial. With 70 sentences, Cohen's $\kappa$ for pairs of annotators was found to be satisfactory (.52, .6 and .43) and motivated us to complete the job on the rest of the sentences. The job was completed upon a compensation of 60€.

### A.1 Task Description
**Approx Duration: 3 hours**

In this annotation trial, you will assess if texts are emotional or neutral.

**Neutral** sentences are those which

1. bear no affective connotation.

**Emotional** sentences are those which either

2. describe an event, a concept or state of affairs to which you would associate an emotion;

3. have an emotion as a central component of their meaning.

Examples of 1. are:

*I am wearing my mask.*
*She answered her phone.*
*A new deal was established between the parties.*
*The elections are over.*

Examples of 2. are:

*I saw my bestfriend.*
*She was being pretty arrogant to me.*
*A war started in Westeros.*
*The king found an old sausage under his bed.*

Examples of 3. are:

*I am so happy to see you.*
*She was bursting with arrogance.*
*And there she was, desperate for her family.*
*I couldn't stand the catering food, bleark!*

### A.2 Guidelines

You will be shown individual sentences and, for each of them, you will answer 3 questions (**Q1**, **Q2**, **Q3**). Go for your immediate reaction to the text – avoid over-assessments.

**Q1** Given a sentence, please ask yourself: **is it emotional (E) or neutral (N)?** Type:

- **N**, if the text does not convey any emotion, like in Examples of 1.;
- **E**, if an emotion could be inferred from the text, like in Examples of 2., or an emotion is a central part of the text, like in Examples of 3.

**Q2** Ask yourself: **how confident am I about my answer to Q1?** Give yourself a rating on a scale from 1 to 3. Indicate if you are

- **1**, not confident at all;
- **2**, confident;
- **3**, sure.

**Q3** This question only applies if you answered 2 or 3 in **Q1**: in case the sentence expresses an emotion, **how strong is such emotion?** Assess the degree of its intensity on a scale from 1 to 3, where

- **1** is mild;
- **2** is intense;
- **3** is very intense.

## B Binary Classification

After data pre-processing step (i.e. exclusion of academic texts and sentences containing words that are masked for copyright reasons), we use a binary classifier as a guide for the selection of sentences from COCA. Using HugginFace[7], we fine-tune the pre-trained BERT base-case model on data for emotion recognition, adding a classification layer that

[7] https://huggingface.co/transformers/

| | P | R | F1 |
|---|---|---|---|
| Emotion | .88 | .84 | .86 |
| Neutral | .89 | .92 | .90 |

Table 3: Binary classification on UED test set.

outputs the labels *emotion* or *neutral*. Data are the resources by Liu et al. (2007), Troiano et al. (2019), Scherer and Wallbott (1997), Alm et al. (2005), Li et al. (2017), Ghazi et al. (2015), Mohammad and Bravo-Marquez (2017), (Mohammad, 2012) and Schuff et al. (2017). We make their format homogeneous with the tool made available by Bostan and Klinger (2018); next, as the labels in the resulting unified emotion data (UED) are not binary, we map *neutral* and *no emotion* instances into *neutral*, and the rest into *emotion*. The total 136891 sentences are then split into train (70%), validation (10%) and test (20%) sets. Classifier's performance is in Table 3.

## C Further Analysis

### C.1 Disagreements

Table 4 reports the distribution of the scores of confidence and intensity for the items where the annotators disagree. This is observed on annotator pairs. A row considers all those items on which either annotators (on the columns) chooses the *emotion* label and the other selects the *neutral* one.

For instance, A1–A2 disagree in total 201 times: on 24 sentences, A1 makes the *emotion* choice, and on 177 sentences it is A2 who picks the class *emotion*. Out of the 24 items, A1 rated 23 as having low intensity and 1 as medium intensity; out of the 177 sentences, 149 are considered of low intensity by A2, 27 as mild, and only 1 as highly intense.

What emerges overall is that people rarely disagree when the *emotion*-leaning annotator has extreme confidence, or perceives very high intensity.

### C.2 Agreements

A manual analysis of the annotations reveals that perfect agreement occurs in the presence of certain patterns. Items unanimously considered emotional often report personal impressions about state of affairs or the speakers' interlocutors (e.g., "*Paris is so sexy*", "*Your expression changed from excited puppy to crestfallen*"), and mostly involve first-hand experiences of the speakers themselves (e.g., "*We'll miss you, but we'll be watching*", "*I'm afraid I don't see anything very beautiful right now*", "*Others helped me and it made a huge difference*"). Instead, sentences that received 3 *neutral* labels seem to be centered on factual statements, like "*Furthermore, the types of materials of manufacture are different*", "*They continue walking*".

One difference between the *emotion* and *neutral* labels is the frequency of agreement, as we found that people concur more on the former – and this invalidates our expectation that, not being given a varied set of affective categories, and not identifying *what* emotion they are judging, people would tend to resort to the neutral choice. Moreover, annotators converge more on one or more on the other label depending on the genre of a text: looking at the distribution of the 304 unanimous *emotion*s (magazine: 28 sentences, blogs: 44, news: 27, tv: 67, fiction: 54, spoken: 39, web: 45) and the 88 *neutral*s (magazine: 18 sentences, blogs: 12, news: 22, tv: 4, fiction: 5, spoken: 12, web: 15), we see that people recognize that affect often manifests itself in fictions, for instance, but is rarer in news – the opposite holds for the neutral expressions.

An obvious strategy to recognize emotions would be to find an emotion name in text. But this is not the case. Sentences that contain emotion words considered less emotionally intense than others: the majority of sentences with CONF3-INT2 contain emotion words (e.g.,"*I was sad to leave.*", while those with CONF3-INT3 are related to extremely negative states of mind (e.g.,"*[...] if I could die and bring her back , I would , but I can't , and I have to deal with that now*").

In Table 5, we report some example sentences on which annotators reached perfect agreement across all confidence-intensity combinations, having choosen the label *emotion*. The sentences are extracted from a number of genres in COCA, and are associated to different scores of intensity (INT) and confidence (CONF). Note that there is no instance that elicited a high intensity evaluation (3) and a low confidence (either 2 or 3) in the annotators. Instances of CONF3 and INT1 show that the correlation between confidence and intensity, though intuitive, has counterexamples.

| | A1 | | | | | | A2 | | | | | | A3 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | CONF | | | INT | | | CONF | | | INT | | | CONF | | | INT | | |
| | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 |
| A1-A2 | 15 | 8 | 1 | 23 | 1 | 0 | 54 | 95 | 28 | 149 | 27 | 1 | – | – | – | – | – | – |
| A2-A3 | – | – | – | – | – | – | 13 | 25 | 6 | 37 | 7 | 0 | 56 | 61 | 3 | 94 | 26 | 0 |
| A3-A1 | 5 | 6 | 0 | 10 | 1 | 0 | – | – | – | – | – | – | 95 | 128 | 17 | 169 | 70 | 1 |

Table 4: Distribution of INT and CONF for disagreements in the EMO task. For a pair of annotators (rows), disagreements are counted when either annotators (i.e., on the columns) chooses the *emotion* class.

| Text Genre | CONF | INT | Text |
|---|---|---|---|
| Magazine | 1 | 1 | I was always a little wary of Arya and Sansa (who also did a little Stoneheart-style vengeance last year) taking on their mother s role . |
| News | 1 | 1 | You can't stress because you just have no idea what 's going to happen. |
| TV | 1 | 1 | So maybe Willy's hanging around the wawa one night , looking for a party...Yeah . |
| TV | 1 | 2 | Mmm , Lordy , Lordy , Lord have mercy . |
| Web | 1 | 2 | Frost is trying to reconcile impulse with a conscience that needs goals and harbors deep regrets. |
| Magazine | 1 | 2 | Nature always solves her own problems ; and we can go far toward solving our own if we will listen to her teachings and consort with those who love her. |
| Fiction | 2 | 1 | "I'm fine ," he replies absently , eyeing the open book . |
| Web | 2 | 1 | My sister likes her map : ) HI Chery : ) lol I'll take'em where I can get'em ... |
| Magazine | 2 | 1 | Whereas coping well means dealing successfully with problems and setbacks, savoring-glorying in what goes right-is an equally crucial emotional competence. |
| Blogs | 2 | 2 | The soldier talks about child detainees . |
| TV | 2 | 2 | We did n't get to bury the others . |
| TV | 3 | 1 | I bruised my lip . |
| Fiction | 3 | 1 | "You have such an interesting life," she said, after a little small talk |
| Magazine | 3 | 1 | "Chalk is unforgiving, " says Oates . |
| News | 3 | 2 | They looked happy, confident . |
| Blogs | 3 | 2 | I am constantly traveling for my job with DISH , and I hate missing all my shows . |
| Blogs | 3 | 2 | I hope I can work through my feelings and keep his friendship in my life. |
| Web | 3 | 3 | I will completely destroy them and make them an object of horror and scorn , and an everlasting ruin. |
| Spoken | 3 | 3 | "We 're very worried ." |
| Spoken | 3 | 3 | If – if I could die and bring her back , I would , but I can't , and I have to deal with that now. |

Table 5: Sentences on which the annotators reached perfect agreement on EMO, CONF, and INT.